DOCUMENT RESUME

ED 191 893 TM 800 517

AUTHOR Kane, Michael T.
TITLE Interpreting Variance Components as Evidence for Reliability and Validity.
PUB DATE Apr 80
NOTE 57p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Attribution Theory; Behavioral Sciences; *Error of Measurement; Mathematical Models; *Measurement; Observation; Physical Sciences; *Sampling; *Test Reliability; *Test Theory; *Test Validity
IDENTIFIERS *Generalizability Theory; Invariance Principle

ABSTRACT

The reliability and validity of measurement is analyzed by a sampling model based on generalizability theory. A model for the relationship between a measurement procedure and an attribute is developed from an analysis of how measurements are used and interpreted in science. The model provides a basis for analyzing the concept of an error of measurement, the distinction between random errors and systematic errors, standardization of measurement procedures, the distinction between reliability and validity, and an analysis of convergent and discriminant validity. Within the sampling model, the concepts of reliability, validity, and import arise naturally as necessary requirements for the results of measurement to be meaningful. This model also provides a basis for examining the relationship between measurement and theory, and reviews some general techniques for developing theory and controlling errors of measurement. (Author/CP)

INTERPRETING VARIANCE COMPONENTS AS EVIDENCE FOR RELIABILITY AND VALIDITY

Michael T. Kane
National League for Nursing

Paper presented at the annual meeting of the American Educational Research Association

Boston, April, 1980

I INTRODUCTION

The technical quality of behavioral measurements is generally evaluated in terms of two properties, reliability and validity. Reliability is associated with the precision of measurement, and reflects the degree of consistency among independent observations. Validity is concerned with the "meaning" of measurement - that is, with the interpretation to be given to the observations.

The aim of this paper is to provide an analysis of the measurement of dispositional attributes. A sampling model for the relationship between a measurement procedure and a dispositional attribute is developed from an analysis of how dispositional terms are used and interpreted in science. The model provides a basis for analyzing the concept of an error of measurement, the distinction between random errors and systematic errors, standardization of measurement procedures, the distinction between reliability and validity, and convergent and discriminant validity. Within this sampling model the concepts of reliability and validity arise naturally as necessary requirements for the results of measurement to be meaningful.

All reliability indices describe the agreement among repeated measurements on the same individuals. The separate measurements for each individual must differ in some of their conditions of observation, and the different reliability coefficients allow different conditions to vary from one set of observations to another. Although they differ in their definitions of error, the different reliability indices all assume that there is a single undifferentiated source of errors.

Generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) provides a multifaceted analysis of the consistency of measurement by relating the variance in observed scores to the sampling of different kinds of conditions of observation, and estimates the relative impact of the various sources of inconsistency. Therefore, generalizability theory provides a general framework in which to examine the dependability of measurements, and it is this framework which is used throughout this paper.

Such a general framework does not exist for the validity of measurement. Criterion validity examines the agreement between observed scores and some external criterion, and typically uses correlation coefficients to yield a single numerical estimate of validity. Content validity examines how well the operations employed in a measurement procedure match the characteristic being measured, and the results of a study of content validity are usually stated as qualitative judgements, rather than as a numerical coefficient.

Construct validity is more general than either content validity or criterion validity. It emphasizes the legitimacy with which various inferences can be drawn on the basis of observed scores, and allows for a wide range of techniques, corresponding to the range of inferences to be drawn. Construct validity may employ the methods of criterion validity or of content validity, but may also use a variety of other techniques.

Unlike reliability, which is defined in terms of agreement among observed scores, validity involves the interpretation of an observed score as representative of some quantity which is not directly observable. Validity requires the assignment of meaning to observed scores.

In introductory textbooks, validity is often equated with the extent to which an observed score measures, "what it is intended to measure." Although this statement is too vague to provide an adequate definition, it does emphasize two important points about validity. First, the looseness of the

statement allows for a very wide range of procedures for evaluating the validity of a measurement procedure, and this is consistent with practice. Second, it suggests the existence of a "real" value for an attribute, without specifying what this "real" value represents. It is often implicitly assumed that the "real" value of the attribute exists somewhere, and that validation requires a comparison, direct or indirect, between this real value and the observed score. Criterion validity tends to encourage this process of reification by introducing the notion of the criterion, which is easily confused with the "real" value of the attribute.

Reliability involves comparisons among observed scores and is intended to indicate the consistency of measurements. Validity seeks to establish an appropriate interpretation for observed scores. Since a high degree of consistency in measuring the wrong attribute is generally seen as being less useful than a lower degree of consistency in measuring the intended attribute, validity is generally considered to be more important than reliability.

Given the great importance assigned to validity, it is surprising that the evidence for the validity of most behavioral measurements is less adequate than the evidence for their reliability. In many cases, evidence for validity is practically nonexistent. Ebel(1961) has aptly described this dilemma:

> "Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development."

This situation has not improved markedly since 1961.

Ebel(1961) also points out that physics, which Campbell(1957) has called, "the science of measurement", does not seem to encounter problems of validation. One reason for this difference between the behavioral sciences and the physical sciences is that much of psychometric theory gives more attention to statistical procedures and assumptions than it gives to the analysis of how measurements are used (Construct validity, as developed by Cronbach and Meehl, 1955, and Cronbach, 1971, is a clear exception to this generalization). In the physical sciences, this situation is reversed; there, the statistical methods used to evaluate measurement procedures are relatively simple, but these methods are closely related to the practice of measurement and its interpretation.

The next two sections are devoted to a discussion of how attributes and measurements of attributes are interpreted. Several simple examples of physical measurement will be introduced in this discussion. These examples are used because the connection between the interpretation of measurements and the indices used to evaluate the accuracy of these measurements is particularly clear in physics.

Generalizability theory (Cronbach et al, 1972) provides the framework and methodology for this paper, and most of the results derived will be stated in terms of variance components. However, the emphasis throughout the paper is on the issues that can be addressed in generalizability theory, rather than on the statistical models. Except for an occassional remark about the severity of some estimation problems, there is no discussion of the complex estimation issues associated with generalizability theory.

OVERVIEW

The analysis presented in this paper is quite long, and, in some ways relatively convoluted. An overview of the main points in the development may therefore provide a useful roadmap.

Section II examines the interpretation given to dispositional attributes. Dispositions can be operationally defined in terms of classes, or universes, of possible observations. The numerical value to be assigned to an attribute is defined as the expected value over this universe, and measurements are interpreted as estimates of this expected value.

Section III analyzes the process of measurement and derives some assumptions that are implicit in the ordinary interpretation of observed scores. The estimates generated by a measurement procedure are based on samples from the universe; consequently, the model for measurements is a sampling model. Estimates of the expected value over a universe, based on different samples, will not generally be equal, and, in order to maintain consistency in the interpretation of measurements, an explicit theory of errors must be introduced. Therefore, the definition of error is based on the interpretation given to the measurements, and errors of measurement are defined by substantive considerations rather than statistical assumptions.

Section IV outlines the terminology and notation of generalizability theory and introduces a sampling model for validity. The validity of measurements of a dispositional attribute is defined in terms of the accuracy with which the observed scores estimate the expected value for the appropriate universe. Where the observed scores are obtained by drawing simple random samples from the appropriate universe, an index of validity can be obtained directly by estimating a generalizability coefficient. In practice, the sampling model for validity becomes quite complicated when it is modified to take account of the sampling designs actually used in measurement procedures.

Section V examines the effects of the standardization of measurement procedures. Standardized measurements involve two kinds of errors: random errors, which vary from one observation to another, and systematic errors, which are constant for a series of measurements. Random errors are related to reliability, and systematic errors are related to validity; this analysis makes it possible to draw a clear distinction between reliability and validity.

Section VI discusses the relationship between the development of theory and the definition of dispositional attributes. A third property of measurements is introduced by defining the concept of import in terms of all of the inferences that can be drawn from an observed score. This section reviews some potentially powerful techniques for developing theory and controlling errors of measurement.

Many of the results derived in this paper are based on rather strong sampling assumptions. In particular, the unbiased estimation of variance components, which are used extensively in this paper, requires random sampling assumptions. In most cases of practical interest, these assumptions are likely to be violated. In section VII, these assumptions, and the robustness of the results derived from these assumptions, are examined. Section VII also presents some concluding comments.

II The Interpretation of Measurable Attributes

Lord and Novick(1968, p.17) define measurement as "a procedure for the assignment of numbers ... to specified properties of experimental units in such a way as to characterize and preserve specified relationships in the behavioral domain". In discussing the methodology of physics, Campbell(1957, p.267) defines measurement as "the process of assigning numbers to represent qualities".

According to Nunally(1967, p.2) "Measurement consists of rules for assigning numbers to objects to represent quantities of attributes." If the word "objects" is interpreted broadly to include persons and groups of persons as well as physical objects and systems, Nunally's definition applies to measurement in both the physical and the behavioral sciences.

Measurement consists of the mapping of objects into real numbers, and establishes a functional relationship between real numbers and the members of some class of objects. Depending on the attribute being considered, the object of measurement may take a variety of forms, including physical objects, persons, pairs of objects or persons, groups, and various complex systems. The rules used to assign the numbers may also vary considerably. However, the process of measurement always involves a mapping of the form:

$$u_o = A(o) \qquad (2.1)$$

where o is an object, A represents the rules used to assign numbers for the attribute, and u_o is the real number assigned to o for the attribute, A.

Note that Eq(2.1) makes a fundamental theoretical commitment, in that it implies that the attribute depends only on the object of measurement and does not depend on any of the conditions that may prevail when the observations are made. For example, the statement that the length of a particular rod is 10 inches can be represented as:

$$10 = L(r)$$

where L represents the procedures used to measure length in inches and r represents the rod. This formulation implies that the length of the rod does not depend, for example, on the location, orientation, or temperature of the rod. The length is also assumed to be independent of the person who carries out the operations represented by L.

Eq(2.1) provides a very general symbolic representation of the process of measurement. However, this definition is not very informative or very useful unless the nature of the objects o and the functions, A, that are involved are well understood. In practice, both the function, A, and the object, o, may be quite complex.

ATTRIBUTES

A measurable attribute can be viewed as a disposition, or a tendency to react in a certain way to some kind of conditions. Dispositions may be qualitative or quantitative. For a qualitative disposition, the object is said to have the attribute if a specific reaction occurs, and is said not to have the attribute if the specific reaction does not occur. A classic example of a qualitative disposition is the property of being a magnet. The typical test condition would consist of placing a small piece of iron near the object being tested. If the iron tends to move toward the object, the object is said to be a magnet, and if the iron shows no tendency to move toward the object, the object is said to be nonmagnetic.

For a quantitative disposition, a number is assigned to the object on the basis of the strength of the reaction to the test conditions. The magnitude of the attribute of being magnetic, or the strength of a magnet, could be defined by how far it moves a piece of iron.

There are basically two ways in which measurable attributes are introduced into science. In the early stages of any science, attributes are developed by quantifying ordinal relationships. Subsequently, other attributes can be derived from empirical laws.

The process of quantifying ordinal properties is one of explication, the transformation of subjective observations into a relatively well defined measureable attribute. Attributes that are defined in this way will be called basic attributes. It is noticed, for example, that some objects are easier to move than others. It is also noticed that this ordering of objects remains the same regardless of where the objects are located, who attempts to move them, or when they are moved. It is convenient, therefore, to think of "resistance to movement" as a property, or attribute, of the objects, and a large class of solid objects can be rank-ordered in terms of this property. Where such an ordinal property exists for a class of objects, numbers can be assigned to all objects in the class, such that the ordering of the numbers corresponds to the ordering of the objects. This assignment of numbers defines an ordinal scale for the attribute.

After some basic attributes are developed, empirical laws that state relationships among those attributes can be developed, and these laws often involve constants that can also be treated as measurable attributes. For example, the measured length, l_r, of a metal rod is found to vary systematically with the temperature, t_r, of the rod. If dt_r is a change in the temperature of a rod and dl_r is the corresponding change in length,

$$dl_r = k_r \, dt_r \, l_r \qquad (2.2)$$

where k_r is a constant, called the coefficient of thermal expansion of the rod. Estimates of k_r are obtained by changing the temperature of the rod, and measuring this change in temperature and the corresponding change in length. An estimate of k_r is then given by the ratio of the change in length to the change in temperature. The operational definition of k_r depends on the definitions of length and temperature, two basic attributes, and on the empirical law in Eq(2.2) which states a relationship between the two basic attributes. Therefore the interpretation of the coefficient of thermal expansion as a measurable attribute is derived from the interpretation of two basic attributes, and the law relating these two attributes.

The value of k_r varies from one rod to another but remains relatively constant from one observation to another on a given rod. It is convenient, therefore, to interpret k_r as a property of the rod by assuming that k_r depends on the rod but not on the conditions prevailing when the rod is observed:

$$K_r = K(r) \qquad (2.3)$$

The assumption that k_r doesn't depend on the conditions of observation is a good approximation over a wide class of observations which is taken to be the universe of generalization. Any observation from this universe of generalization could be used to estimate k_r.

The interpretation of numbers as the values of an attribute depends on empirical laws that state that different observations on any pair of objects generally rank-order the objects in the same way. Therefore the results of

any set of observations provide information about a much wider class of observations that could have been made. It is this generalization from particular observations to a universe of observations that provides the meaning of an attribute and that makes measurements of the attribute useful.

OPERATIONAL DEFINITIONS

The rules that are used to assign a value to an attribute are usually called operational definitions(Bridgman, 1927). The rules are operational in the sense that they are stated in terms of the operations performed in measuring the attribute. The rules are said to be definitions because they provide most of the meaning of the attribute; that is, they provide a basis for interpreting the numbers assigned as values of the attribute. (Ennis, 1973, Hempel, 1960, and Carnap, 1953, provide analyses of the use of operational definitions in science)

Operational definitions generally include two kinds of rules, structural rules and selection rules. The structural rules specify the kind of observations that are to be used, and the way in which numbers are to be derived from these observations. Thus, psychologists arrange stimulus situations that are likely to elicit the type of behavior that they wish to study. In the absence of such standardization of some characteristics of the observations, it would be very difficult to provide rules for the assignment of numbers. The structural rules may be more or less elaborate. A rule for measuring the length of a rod, for example, might require that an observer align the zero mark of a tape measure with one end of the rod, and record the number on the tape measure that coincides with the other end of the rod. More detailed rules for measuring length are discussed by Campbell(1957), and others, but all of these rules leave some issues open. For example, what kind of tape measure is to be used? Could a slightly bent metal tape measure be used, and would an observer with astigmatism be acceptable?

Such questions lead to the development of selection rules. The selection rules specify the range of conditions that may be tolerated for the various characteristics of the observations. Some of these characteristics may be fixed; in the example above, one end of the rod must be aligned with the zero mark of the tape measure. Other characteristics are specified in terms of ranges for continuous variables; for example, the temperature is to be between 15°C and 25°C. It is assumed that the characteristics not mentioned in the structural rules need not be controlled at all.

As the example of length indicates, operational definitions do not specify particular observations; they specify classes of observations. The rules for measuring length that were sketched above could be made more precise and more complete by specifying a particular tape measure, a particular temperature, etc., but it would be impossible to specify all of the characteristics that might influence an observation. Furthermore, it would be self-defeating to make the rules too specific because this would limit the usefulness of the concept of length. Ceteris paribus, scientists prefer to use concepts which are as general as possible.

Operational definitions are designed to achieve this generality of application, while providing a clear specification of the class of observations allowed. The fact that operational definitions specify classes of observations rather than specific observations does not necessarily imply any lack of precision in the definition since classes can be defined precisely. In practice, however, these classes are always defined somewhat ambiguously. If the lighting in the room and the vision of the observer are discussed at all for measurements of length, the requirement is likely to be that they be "normal' or "within normal limits." Most of the characteristics of an observation are ignored unless there is some reason to believe that a

particular characteristic is extreme enough to have a serious effect on the observation. This ambiguity is recognized and tolerated because it makes general laws possible (Toulmin,1953).

THE OBJECT OF MEASUREMENT

The object, or unit, to which a number is assigned by measurement is called the object of measurement. The number representing the attribute is not assigned to an observation. The operational definition of an attribute specifies a class of observations and no one of these observations has special significance.

A measurement assigns a number to some object of measurement which is involved in, and partially defines a number of observations. The purpose of measurement is to map objects of measurement into real numbers. The number assigned to each object is intended to represent the magnitude of the attribute for the object of measurement. A particular observation can provide information about different kinds of objects of measurement, and if the measurement is to be interpreted unambiguously, it is necessary to clearly identify the object of measurement. Cardinet, Tourneu, and Allal(1976) have discussed the kinds of units which may serve as objects of measurement.

In a study of anxiety an observation might consist of the response of a person to some stimulus in a particular context. For such observations, the person is usually taken as the object of measurement. However, for researchers seeking to determine how anxiety provoking stimuli or contexts are, the objects of measurement would be stimuli or contexts, respectively, instead of the person.

More complicated objects of measurement can also be considered. The differential impact of stimuli on different persons could be investigated by taking person-stimulus pairs as the objects of measurement. A researcher who is investigating the interactions between persons and contexts might take person-context pairs as the objects of measurement.

Although the observation has a single assigned value, this value may be given various interpretations depending on the definition of the object of measurement. The specification of the object of measurement is a conceptual issue, and is not uniquely determined by the nature of the observations that are made. As the examples above illustrate, a single observation can provide information about a variety of objects of measurement. Similarly, many different observations may be used to measure a particular attribute for an object of measurement.

The distinction that is often drawn in psychology, between a state and a trait depends on a distinction between different kinds of objects of measurement. If the object of measurement is taken to be a person in a particular context, then the attribute being measured is a state variable, which is assumed to be a function of both the person, and the context or time. It is, therefore, expected that the value associated with a state variable will change as the context changes over time. For a trait, however, the object of measurement is the person, and the value of the trait variable is assumed to be independent of time. It is recognized of course that the behaviors associated with the trait variable will be exhibited to different degrees in different contexts, but this is true of all dispositional variables. For a trait variable, changes in the observed variable over time are taken as errors of measurement; for a state variable such differences are accounted for by differences in the value of the state variable.

A similar distinction is made in physics between mass and weight. Mass is defined to be an attribute of a physical object, while weight is defined to depend on the physical object and the location of the object in a gravitational field.

In the physical sciences, the object of measurement to which attributes are assigned are specified explicitly. In their introductory treatment of mechanics, Corben & Stehl(1960) state the following assumptions:

> A particle is described when its position in space is given and when the values of certain parameters such as mass, electric charge, and magnetic moment are given. By our definition of a particle, these parameters must have constant values because they describe the internal constitution of the particle. If these parameters do vary with time, we are not dealing with a simple particle. The position of a particle may, of course, vary with time.(p.6)

Therefore the mass, charge, and magnetic moment are to be treated as trait variables, with particles as their objects of measurement. Position, however, is to be treated as a state variable with particle-time combinations as its objects of measurement.

It is sometimes claimed that operationally defined attributes are valid by definition. It is maintained that the operations used to measure the attribute define the attribute, and the results of these operations are, by definition, the values of the attribute. According to this view no interpretation is to be given to the numbers assigned to objects beyond the fact that they result from the particular set of operations. Therefore there is no inference from the results of the operations to a wider class of observations and no need to check on the accuracy of inference.

However, in practice, the operational definitions of even the most narrowly defined attributes involve classes of observations rather than particular observations. No operational definition that is of any significance in science specifies a particular observer(John Jones), particular equipment (voltmeter #6) and particular time and place. Although restrictions may be placed on the qualifications of observers and on the type of equipment used, these restrictions define classes of observations rather than particular observations. If the results of a particular observation could not be used to draw inferences about similar observations, these results would be of little interest.

Unlike observations, attributes are not unique to a particular combination of time, place, observer, etc. Attributes have a generality that is not associated with observations, and to assign a value to an attribute is to make a claim about an infinite class of observations. Since few of these observations will actually be made for any object of measurement, all attributes are, in a sense, theoretical constructs.

Attributes are "constructed" by specifying classes of observations. Measurements of attributes are based on samples from these universes. In order to interpret the results of measurement as the value of an attribute for the object of measurement, we must generalize from a sample of observations to a universe of observations, and these generalizations are inductive inferences. A central concern of a theory of measurement is the justification of such inferences.

III Measurement of Disp Attributes

The measurement of a dispositional attribute involves a generalization, or an inductive inference, from a particular observation to the universe of observations defining the attribute. These inductive inferences require justification, and it is the task of measurement theory to provide an analysis of the kind of justification that is required.

THE USE OF INVARIANCE PROPERTIES AS INFERENCE TICKETS

The justification for scientific inference is generally provided by appeal to scientific laws. Hempel(1965) has discussed the use of laws as a basis for scientific inference in considerable detail, and Toulmin(1953), who sees the inferences derivable from a set of laws as defining the content of the laws, has suggested the term "inference tickets" for scientific laws when they are used in this way.

The type of law that is needed to justify the interpretation of observations as measurements is an invariance property. An invariance property states that the results of a certain kind of observation do not depend on some of the particular conditions of observation. Invariance properties are necessarily involved in measurement because measurement assigns a value to an attribute of some object of measurement on the basis of some set of observations.

A complete description of an observation would require exhaustive specification of all of the conditions under which the observation is made. Since the operational definition of an attribute specifies only some of the conditions of the observations, thereby allowing the other characteristics to vary, it describes a class of observations rather than a single observation. The attribute is identified with this class of observations and not with any particular observation. For any attribute, the conditions of observation are limited by the selection rules but are not uniquely specified. Any one of this class of observations could be used to assign a value for the attribute to the object of measurement. The observed score, X_{oi}, for an observation is the real number assigned to the observation by the structural rules for the attribute.

A different but equally legitimate observed score, $X_{oi'}$, could be obtained by changing the conditions of observation in accordance with the selection rules.

When an observed score is interpreted as a measurement, it is assumed, implicity or explicitly, that this observed score can be taken as the value of the attribute for the OM. Since the observed scores for different observations may be assigned to the value of the same attribute for an OM, the following relationship must hold at least approximately in order to maintain consistency:

$$X_{oi} = X_{oi'} \qquad (3.1)$$

where i and i' represent any two observations for the object, o, that meet the specifications of the operational definition for the attribute. That is, the observed scores must be invariant over the universe of observations defining the attribute. Since the two quantities in Eq(3.1) are based on observations, this assertion is testable for any pair of observations, and Eq(3.1) is an empirical law.

Eq(3.1) is an invariance property stating that the observed scores are invariant over the universe defining the attribute. For a given object of measurement, o, all observations included in the definition of the attribute

should assign the same value to the object of measurement. If all observations do assign the same value to the OM, this value is taken to be the value of the attribute, u_o, for the object of measurement, o.

Note that the invariance properties required for the measurement of an attribute depend on the definition of the universe for the attribute being measured. If the two quantities in Eq(3.1) were not taken as measurements of the same attribute for the same OM, there would be no reason to require that they should have the same value.

Considering again the example discussed earlier, if anxiety is interpreted as a trait, the situations in which observations are made are conditions of observation, and invariance over these situations is assumed. If anxiety is interpreted as a state, the objects of measurement are persons in situations, and changes in the value of the observed score as a function of the situation are consistent with this interpretation.

Invariance properties are involved in measurement because they justify inferences from samples of observations to a universe of observations. If all of the observations in the universe give the same result for any object of measurement, then any one of these observations provides complete information about the universe. If Eq(3.1) holds for all pairs of observations defining an attribute, it provides the necessary justification for inferences from observed scores to universe scores. To the extent that observations fail to satisfy Eq(3.1), such inferences are not justified. Therefore, the invariance property in Eq(3.1) is necessary for the interpretation of observations as measurements of dispositional attributes.

If the observations in the UG for each object of measurement do not all yield the same observed scores a variety of values are assigned to a single quantity, the universe score. Therefore, violations of the invariance property in Eq(3.1) imply inconsistency in inferences from observed scores to universe scores. If the magnitude of the discrepancies is generally small, it may be possible to ignore them, and to treat Eq(3.1) as an approximation, resulting in what Suppes(1974) has called a deterministic theory without a theory of error. The alternative is to introduce an explicit theory of errors.

ERRORS OF MEASUREMENT

In order to develop an index for the accuracy of this approximation, and therefore of the dependability of inferences from observed scores to universe scores, the concept of an error of measurement is introduced. The result of any observation on an object, o, is taken to be the sum of the "true" value of the attribute, u_o, plus an error of measurement, e_{oi}.

$$X_{oi} = u_o + e_{oi} \qquad (3.2)$$

Since neither the "true" value nor the error of measurement is directly observable, Eq(3.2) is not a testable hypothesis; rather, it is a definition of the variable, e_{oi}. For any observation and any value of u_o, the value for e_{oi} can be chosen so that the two sides of Eq(3.2) are equal, and therefore Eq(3.2) is a tautology.

However the values assigned to the error term in the formulation presented above are not arbitrary. Given the UG for an attribute and any value for u_o, the magnitudes of the errors are determined empirically. If the observations on a given object of measurement vary widely, the magnitudes assigned to the errors must be large, and if the observations have approximately the same value, the errors can be taken to be small. Measurements with small errors of measurement are, of course, generally

preferred to measurements involving large errors of measurement, and Eq(3.2) provides the basis for a relative criterion for the dependability of measurement.

Although this development has assumed that u_o is a constant for any object of measurement, the value of this constant has not been specified. In both the physical and behavioral sciences, u_o is generally equated with the mean over all observations allowed by the operational definition of the attribute:

$$u_o = \underset{i}{E}\,(X_{oi}) \tag{3.3}$$

Eq(3.3) determines a unique value of the attribute for each OM. This choice is a convention, which is both convenient and plausible, but it is a convention. The most compelling reason for this convention is that it minimizes the mean-square error.

With the value of an attribute defined by Eq(3.3), it is easy to show that the expected value of the errors of measurement, as defined by Eq(3.2), is zero for each object of measurement.

$$\underset{i}{E}\,(e_{oi}) = \underset{i}{E}X_{oi} - u_o = 0 \tag{3.4}$$

The error variance for each object of measurement is given by:

$$\sigma^2_o(e_{oi}) = \underset{i}{E}\,(e^2_{oi}) \tag{3.5}$$

The error variance in Eq(3.5) is a measure of the dispersion in estimates of the universe score for each object of measurement. Since u_o is a constant for each OM, the error variance in Eq(3.5) is equal to the observed score variance for the OM. Where they can be estimated, the error variances in Eq(3.5) are very useful because they provide an indication of the accuracy of inferences from observed scores to universe scores, for each OM. In the physical sciences, the precision of measurement is often reported for each OM in terms of the square root of Eq(3.5). However, the direct estimation of this error variance requires repeated observations on each OM, and this is often not practical for behavioral measurements.

A more easily estimated parameter is the expected error variance over the population, as given by:

$$\sigma^2(e_{oi}) = \underset{oi}{E}\,(e^2_{oi}) = \underset{o}{E}\,\sigma^2_o(e_{oi}) \tag{3.6}$$

Eq(3.6) provides an indication of the accuracy of inferences from observed scores to universe scores, averaged over all objects of measurement. Although Eq(3.6) doesn't provide seperate indices of the accuracy of measurement for each OM, it does provide a useful overall index of accuracy for the population. The average error variance can be estimated with pairs of observations on objects of measurement.

Using Eq(3.4), the covariance between universe scores and errors of measurement, can be shown to be equal to zero:

$$\underset{oi}{E}\,(u_o - u)(e_{oi}) = \underset{o}{E}(u_o - u)\underset{i}{E}(e_{oi}) = 0 \tag{3.7}$$

where u is the expected universe score over the population, and the covariance is taken over the population and over the UG. Since the errors of measurement are independent of the universe scores, the observed score variance can be partitioned as:

$$\sigma^2(X_{oi}) = \sigma^2(u_o) + \sigma^2(e_{oi}) \qquad (3.8)$$

where $\sigma^2(u_o)$ is the variance in the universe scores over the population and $\sigma^2(e_{oi})$ is the error variance taken over the population and the UG.

Although scientists would prefer to recognize the presence of some error of measurement rather than add complexity to theories, they still seek to minimize the magnitude of such errors. This is usually done by standardizing the conditions of observations in measurement and by basing each measurement on more than one observation. The choice of the mean over the universe of generalization as the value of the attribute is consistent with this tendency to try to minimize errors of measurement.

ERRORS OF MEASUREMENT AS CONSTRUCTS

In the absence of any assumptions about the object of measurement, the concept of an error of measurement is unnecessary. If we restrict our attention to observations, there is no reason to reject the hypothesis that every observation is perfectly accurate. If measurement assigned numbers to observations, therefore, it would not be necessary to introduce the concept of an error of measurement.

Suppose, for example, that two observers put mercury thermometers into the same glass of water, at the same time. Suppose further that one of the observers records that the mercury rises to mark labeled 60 and the other observer records that the mercury rises to a mark labeled 58. These two observations differ in several ways. If the two numbers, 60 and 58, are assigned to the observations, there is no reason to assume that either observation should be said to contain any error. The two observations occurred as they occurred. The concept of an error of measurement arises only when attention is shifted from observation to measurement. The assumptions about invariance properties that are involved in the measurement of any attribute force us to introduce the concept of an error of measurement.

The usual analysis of the example given above takes temperature to be the attribute, and the glass of water at a particular time to be the object of measurement. The temperature is assumed to be a function of the water and the time. This implies that the two observations described above should agree with each other, as indicated in Eq(3.1). That is, temperature is assumed to be invariant over thermometers and over observers.

However, any two observations on the same object of measurement will, in general, produce different numerical results. Since measurement is intended to map each object into one real number, theory must be adjusted in one of two ways. One approach is to redefine the objects of measurement so that the measurements which disagree with each other involve different objects of measurement. In the example given above, the object of measurement could be redefined to be a small volume of water in the glass, at a given time as would be the case in investigations of thermal diffusion. Since the two thermometers must be at different positions in the water, the differences between the two measurements can be explained by the fact that they apply to different OMs. This approach resolves the inconsistency between assumptions and observations, but it does so at the cost of greatly increasing the number of OMs to be considered.

An alternative approach leaves the definition of the object of measurement unchanged, but introduces an explicit theory of errors. It is thereby recognized that the observations used in measurement depend on the conditions of observation as well as the object of measurement.

RELIABILITY COEFFICIENTS

For many applications, the expected error variance in Eq(3.6) is not a very good index for the dependability of measurement. The magnitude of the error variance can be changed simply by changing the scale (e.g. inches to feet or meters), and the evaluation of a measurement procedure should not depend on such an arbitrary choice. Therefore, it is not the absolute magnitude of the error variance that is significant, but the magnitude relative to the degree of precision needed for some purpose.

The degree of precision, or tolerance, required of measurements varies from one area of science to another. The astronomer who is measuring the distances between stars can tolerate errors of thousands of kilometers, while a crystallographer might consider an error of a thousanth of a centimeter to be unacceptable. Between these two extremes, lies a continuum of possible tolerances, including those of the engineer who wants the separate parts of a bridge to fit together. If the error variance in Eq(3.6) is based on the same scale as the tolerance, the dependability of measurement can be evaluated directly by comparing the estimated magnitude of the error to the tolerance. This is the usual method for evaluating the precision of measurement in physics, and since the tolerances in a particular area of investigation are usually well known, it is common practice to report the square root of Eq(3.5) or Eq(3.6) as an index of the precision of measurement.

The practice of reporting the relative magnitude of errors of measurement is sufficiently general in the physical sciences, as to be introduced in the first chapter of an introductory textbook (PSC,1968,p14):

> "If a surveyor measures a distance with great care he might get 100.132 meters ± 0.3 cm. His work is a great deal more accurate than that done when the width of a book page is measured to the nearest millimeter with a ruler, even though his error is something like three times as big as what anyone would perhaps make on the page in ten seconds' work. This sometimes finds expression in another way when the estimated spread of measurements, the tolerance, is stated, using decimal fractions, or percentage. Thus the surveyor would say his length was 100.132 meters ± 0.003%, while the page is ust 20.1 cm. ± 0.5%."

The emphasis on stating the magnitude of the errors of measurement in relative terms has been even more pronounced in the social sciences (Lord and Novick, 1968, p.252):

> "...the effectiveness of a test as a measuring instrument usually does not depend merely on the standard error of measurement, but rather on the ratio of the standard error of measurement to the standard deviation of observed scores in the group. The more discriminating the test items, the larger will be the standard deviation of observed scores, other things being equal; and hence, the less will be the danger that true differences will be swamped by random errors of measurement and lost to view."

A suitable index for the relative magnitude of errors of measurement is suggested by the relationship between dispositional attributes and the rank ordering of the properties of observations. As a minimal requirement, the errors should not be so large as to cause significant fluctuations in the

ranks assigned to OMs from one set of observations to another. If the universe scores for two objects of measurement are u_0 and $u_{0'}$, errors of measurement which are less than the absolute value of $(u_0-u_{0'})/2$ will not distort the ordering of the observed scores for these two objects. In comparing these two objects of measurement, therefore, an error variance which is smaller than $(u_0-u_{0'})^2$ can be considered a relatively small error variance. (see Cronbach and Gleser, 1964, for more detailed analysis of signal-to-noise ratios)

A more direct way of evaluating the consistency of the ranking of objects of measurement from one set of observations to another is to estimate the correlation between observed scores based on independently sampled observations. Correlations indicate the degree of linear relationship between two variables, but, in the absence of very serious departures from linearity, a correlation coefficient depends mostly on the consistency of the rankings from one variable to the other. Therefore, correlations are useful statistics for evaluating the consistency of the rankings of observed scores.

If the only purpose of measurement were to reflect the rank ordering of objects on some attribute, rank order statistics would be more appropriate than correlation coefficients in evaluating measurements. However, measurement provides an explication of ordinal relationships rather than just representing this ordinal relationship. Measurements are intended to assign a number on an interval scale to each OM to represent the value of an attribute. Correlation coefficients are appropriate indices for the precision of such numerical assignments, while rank-order statistics would not be appropriate for this purpose. Therefore correlation coefficients and indices that are closely related to correlation coefficients (i.e. generalizability coefficients) have been widely used in evaluating the dependability of measurements. In particular, correlation coefficents constitute the basic mathematical machinery in classical test theory.

THE ROLE OF THEORY

This analysis of measurement errors depends on the fact that certain assumptions are made about measurable attributes. In particular it is assumed that attributes are to be applied to specific kinds of objects of measurement, and that certain invariance properties will hold. These assumptions are theoretical in the sense that they form a connected body, or network, of general laws. The criterion in Eq(3.3) is a theoretical ideal which is never achieved in practice. Errors of measurement may be viewed as concessions to the brute fact that the world of observations is not as neat and orderly as we might like it to be.

Postulating the existence of errors of measurement makes it possible to minimize the number of objects of measurement that need to be considered, and therefore to simplify both descriptions of phenomena and the theories designed to explain phenomena. The resulting gain in conceptual clarity is usually worth the loss of precision involved in relegating the effects of some conditions of observation to error.

The introduction of an explicit theory of errors represents a decision not to study some kinds of phenomena. In the examples discussed above, the decision to attribute the difference between the two thermometer readings, 58 and 60, to errors of measurement is essentially a decision not to investigate temperature variations within the liquid; this decision, which is not dictated by empirical findings, reflects a choice among several possible research strategies.

The designation of certain sources of variance as errors of measurement is a conceptual choice rather than an empirical finding. Errors of measurements provide a way of handling observational variations that are not

to be given an explicit description or explanation at a particular stage in the development of a science. In order to make its task more managable, every science tends to restrict the phenomena that it treats explicitly. As the science develops, it may be able to analyze phenomena that had earlier been relegated to error variance; this decreases the error variance, and enlarges the sphere of phenomena treated by the science, but there is always some variation which is intentionally left unexplained.

The specification of the attributes in an area of science and of objects of measurement determines how observations are described and organized, and this influences the kinds of questions addressed by the science, i.e. the paradigm for the science. A change in the definitions of attributes and objects of measurement, which is equivalent to a change in the definition of error, represents a shift in the way that phenomena are perceived and described. If the attributes that are redefined are fundamental, the resulting changes may be significant enough to be called a scientific revolution (Kuhn, 1970).

For example, the changes introduced into physics by the special theory of relativity are basically changes in the concepts of length and time; specifically, they are changes in the set of invariance properties associated with length and time. In classical mechanics, length and time are assumed to be invariant with respect to the observer; in the theory of relativity, this invariance property is rejected, and the object of measurement is redefined to include the observer (more precisely, the observer's frame of reference). Special relativity had a revolutionary impact on physics because it modified the fundamental concepts of length and time; analagous changes in less important concepts would have had a much smaller impact. (See Frank, 1953, for a very lucid discussion of this point)

Although the formulation presented here assumes the existence of some set of basic assumptions, these assumptions do not necessarily include a model of underlying constructs or processes. Attributes are treated as dispositions (see Carnap, 1953), and there is no ontological commitment to attributes as things. Attributes are defined in terms of universes of observations, and the assumptions specify the syntax and semantics of the descriptive language of some part of science.

The existence of errors of measurement, therefore, depends on theoretical assumptions about attributes, and the operational definition of errors of measurement depends on the definitions of the attributes and objects of measurement. In particular, the definition of the objects of measurement determines whether the observed differences among observations are to be interpreted as errors of measurement, or as differences in the attribute for different objects of measurement.

IV Generalizability Theory as the Basis for a Sampling Model for Validity

This discussion of generalizability theory is necessarily only a brief outline. A thorough presentation of generalizability theory can be found in The Dependability of Behavioral Measurements (Cronbach et.al., 1972). Introductions to some of the basic ideas in generalizability theory are found in Lindquist (1953) and in Brennan and Kane (1979).

GENERALIZABILITY THEORY

The purpose of both reliability theory and generalizability theory is to characterize the dependability of measurements. Unlike reliability theory, which treats errors of measurement as arising from a single source and uses correlation coefficients as indices of reliability, generalizability theory recognizes the existence of multiple sources of error, and uses a variety of multifaceted designs to estimate the variance components for different sources of error.

In generalizability theory, any observation on an objects of measurement is assumed to be sampled from a universe of observations. The observations in this universe are characterized by the conditions under which they are made, and the set of all conditions of a particular type is called a facet. For example, if persons are the objects of measurements and the dispositional attribute is a state variable, the universe could include an item facet, an occasion facet, and perhaps a rater facet.

Cronbach et.al.(1972, p.20) draw a distinction between G studies, or generalizability studies, which examine the dependability of measurement procedures, and D studies, or decision studies, which provide the data for substantive decisions. In this paper, the term, "measurement procedure", will often be used in place of the term "D study". A measurement procedure incorporates a sampling design for obtaining observations on objects of measurement that is used over a number of separate studies. The term "D study", suggests that the sampling design for measurements of an attribute is likely to change from one study to another. Although the possibility of such changes in D-study design is explicitly considered at several places in this paper, much of the discussion will emphasize the effects of standardization of the conditions of observations. Measurement procedure is a more descriptive term than D study, when some facets are standardized over all observations.

The distinction drawn between a universe of admissible observations and a universe of generalization (Cronbach et al, 1972, p.20) is based on the distinction between G studies and D studies. In conducting a G study, certain facets are investigated, and a certain range of conditions is considered with respect to each facet. The facets investigated in the G study define a universe of admissible observations. In interpreting the observations in a D study as measurements of an attribute, inferences are drawn to the universe of observations that provides an operational definition of the attribute. In generalizability theory, this universe is called the universe of generalization for the attribute.

The universe of admissible observations is associated with estimation, and indicates the facets for which variance components have been estimated in G studies. Since estimation issues are generally not addressed in this paper, few references will be made to the universe of admissible observations. The concept of a universe of generalization, which defines an attribute will be used extensively in paper. In addition, another universe, not discussed by Cronbach et.al.(1972) will later be introduced in order to describe a measurement procedure.

The purpose of the G study is to estimate components of variance, which may then be used to evaluate the dependability of inferences to the universe of generalization. If the components of variance estimated in a G study are to provide the information needed for evaluating the D study, they must provide estimated variances for sources of error in the D study. Therefore, the universe of admissable observations must include the universe of generalization. G studies are most useful when they employ crossed designs and large sample sizes to provide stable estimates of as many variance components as possible. For any measurement procedure, there are many facets that might be considered, but variance components for only a few of these can be independently estimated in any G study. Therefore, several G studies may be required to adequately evaluate the dependability of a measurement procedure.

The universe score, the expected value over the universe of generalization, is stipulated to be the value of the attribute for any objects of measurement. Universe scores are not directly observable, but can be estimated by the mean over a sample of observations; that is, for each objects of measurement, the observed score is used as an estimate of the universe score. In generalizability theory, then, questions about the reliability of a measurement procedure are replaced by questions about the generalizability of observed scores, and the dependability of such generalizations is described by a generalizability coefficient.

Cronbach et.al.(1972, p.97) define the coefficient of generalizability for an attribute and a D study as the ratio of the universe score variance to the expected observed score variance for the D study design. The universe score variance in a generalizability coefficient replaces the true score variance of classical test theory, and the expected observed score variance replaces the observed score variance of classical test theory.

Cronbach et al.(1972, p.98) discuss two interpretations of the generalizability coefficient for a D study which samples from the intended universe of generalization. First, the generalizability coefficient is approximately equal to the correlation between observed scores for two independent random samples of observations from the universe of generalization. Second, the generalizability coefficient is approximately equal to the expected value of the squared correlation between the observed score and the universe score.

A LINEAR MODEL

Generalizability theory allows for the use of a variety of linear models in interpreting the results of both G studies and D studies, depending on the design of the study.

The universe of generalization typically involves a large number of facets, and in principle the model for observed scores could explicitly include any number of these facets. For the sake of simplicity, however, a simple one-facet model with replications will be used as a basis for discussion throughout this paper. In this simple model, one facet is considered explicitly; all other facets in the universe of generalization are assumed to be sampled randomly and independently, and are subsumed under a single replication facet. The observed scores for all observations in the universe of generalization are represented by the linear model:

$$X_{oir} = u + a_o + a_i + a_{oi} + a_r \qquad (4.1)$$

where

u is the grand mean

a_o is the main effect for the object of measurement, o

a_i is the main effect for the i facet

a_{oi} is the oi interaction

a_r is the replication effect

The linear model in Eq(4.1) represents the observed scores in the universe of generalization; it is not intended to represent the sampling design for any particular G study or D study. The universe of generalization defines an attribute for a population.

It is assumed that the i facet is crossed with objects of measurement in the universe of generalization; that is, in the universe of generalization, there is an observed score for each possible combination of an objects of measurement and a condition from the i facet. This does not necessarily imply that D studies or G studies associated with measurements of the attribute will employ crossed designs. Each effect in the model is assumed to be uncorrelated with every other effect. In addition, in order to make the estimates of the effects unique, the expected value of each effect over any of its subscripts is set equal to zero.

Eq(4.1) is essentially a generalization of Eq(3.2). The main difference between the classical test theory model in Eq(3.2) and the linear model in Eq(4.1) is that the classical test theory model assumes the existence of only two sources of variance in the observed score, while the model in Eq(4.1) explicitly considers four sources of variance. The general linear model can be formulated to include as many sources of variance as necessary, and can be made to reflect the design under which the observations are made.

The model in Eq(4.1) includes two facets, labeled i and r, and for each of these facets, there is a universe of conditions from which the conditions in a particular study may be drawn. These universes may be either finite or infinite. Although the consideration of finite universes does not pose a fundamental problem for generalizability theory, it would complicate the discussion, and for the sake of simplicity, it is assumed in this paper that the universe of conditions for each facet is infinite.

From a G study in which conditions of the i facet are crossed with objects of measurement, o, and replications are nested within these oi combinations, four components of variance can be independently estimated. The variance for the four random effects in Eq(4.1) are designated as $\sigma^2(o)$, $\sigma^2(i)$, $\sigma^2(oi)$, and $\sigma^2(r)$.

In a D study, the observed scores for objects of measurements are usually based on the sum or average taken over a sample of observations, and capital letters will be used to designate the average value of an effect over a sample of observations. The variance component for the average value of the main effect, a_i, over a sample of n_i conditions of the i facet is given by

$$\sigma^2(I) = \sigma^2(i)/n_i, \tag{4.2a}$$

Similarly, the variance component for the average value of the oi interaction, over samples of n_i conditions of the i facet, is

$$\sigma^2(oI) = \sigma^2(oi)/n_i, \tag{4.2b}$$

and the variance component for the average value of the replication effect over samples of n_r replications on each of n_i conditions of the i facet, is

$$\sigma^2(R) = \sigma^2(r)/n_i n_r \qquad (4.2c)$$

The fact that the variance components in Eq(4.2) are divided by sample sizes is a reflection of the general statistical principle that sampling variances for means over random samples are equal to the sampling variances for single observations, divided by the number of obsevations in the sample. The relationships listed in Eq(4.2) can be used to estimate variance components for D studies involving any number of conditions of the i facet and any number of replications, once the required random effects variance components are estimated in G studies.

General procedures for the estimation of variance components from computed mean squares are discussed by Cronfield and Tukey (1956), Cronbach et al (1972), Millman and Glass(1967), Brennan(1977), and by most standard textbooks on experimental design (e.g., Kirk, 1968; Winer, 1971).

MEASUREMENT PROCEDURES BASED ON RANDOM SAMPLING FROM THE UG

Although measurement procedures that use random sampling from the universe of generalization are seldom used in practice, it is convenient to start with this oversimplified assumption. Letting capital letters designate effects for samples of conditions, the observed scores for a D study with i nested within o (a separate sample of conditions of the i facet is drawn for each objects of measurement) can be represented as:

$$X_{oIR} = u + a_o + a_I + a_{oI} + a_R \qquad (4.3)$$

where o represents the object of measurement, I indicates a sample of n_i conditions from the i facet, and R indicates a sample of n_r replications for each condition of the i facet. Again, the replication index represents the effect of all facets other than the i facet, and conditions from these facets are assumed to be sampled randomly and independently for each observation. Since the effects in Eq(4.3) are assumed to be independent of each other, the expected observed score variance over the population and over the universe of generalization is:

$$\begin{aligned}\sigma^2(X) &= \sigma^2(o) + \sigma^2(oI) + \sigma^2(I) + \sigma^2(R) \\ &= \sigma^2(o) + \sigma^2(oi)/n_i + \sigma^2(i)/n_i + \sigma^2(r)/n_i n_r\end{aligned} \qquad (4.4)$$

The universe score u_o, for the object of measurement, o, is given by

$$u_o = E_{IR}(X_{oIR}) = u + a_o \qquad (4.5)$$

The universe score variance is given by the variance of the main effect for objects of measurement, a_o:

$$E_o(u_o - u)^2 = \sigma^2(o) \qquad (4.6)$$

Where observations are randomly sampled from the universe of generalization for each objects of measurement, the expected value of the observed score over repeated applications of the measurement procedure is equal to the universe score, and the observed score is an unbiased estimate of the universe score.

In analyzing errors of measurement, Cronbach et.al.(1972, p.76) distinguishes between the error in point estimates of universe scores (which they designate by a capital delta and is represented here by the symbol D) and the error in estimates of the universe score expressed in deviation form (which they designate by a small delta and is represented here by the symbol d). Cronbach et.al.(1972) also discusses a third kind of error, which is based on regression estimates, but is not used in this paper.

The error of measurement for a point estimate of u_o, based on X_{oIR}, is,

$$D_{oIR} = X_{oIR} - u_o$$

$$= a_I + a_{oI} + a_R \qquad (4.7)$$

Since I and R are randomly sampled for each observation on the objects of measurement, the expected value of D_{oIR} over the universe of generalization is zero, and the observed score is an unbiased estimate of the universe score for o. The expected value of the squared error, taken over all instances of the procedure is equal to the average error variance within the objects of measurement, and is given by:

$$E_{IR}E(D^2_{oIR}) = \sigma^2(I) + \sigma^2(oI) + \sigma^2(R) \qquad (4.8)$$

If conditions of the i facet are sampled independently for each observation, the expected value of X_{oIR} over an infinite population of objects of measurement is also an expected value over the universe of generalization and is equal to the grand mean, u. Therefore, if both I and R are nested within o, the error in estimating universe scores relative to the population mean is:

$$d_{oIR} = (X_{oIR} - u) - (u_o - u)$$

$$= a_I + a_{oI} + a_R \qquad (4.9)$$

Since I and R are independently sampled for each observation of o, the expected value of d_{oIR} over the universe of generalization is also zero. The expected value of the squared error, d^2_{oIR}, is equal to the expected value of the squared error, D^2_{oIR}, as given in Eq(4.8).

$$E_{IR}E(d^2_{oIR}) = \sigma^2(I) + \sigma^2(oI) + \sigma^2(R) \qquad (4.10)$$

The covariance, taken over the population, of the errors, D_{oIR}, on two administrations of the measurement procedure is given by:

$$\mathrm{cov}(D_{oIR}, D_{oI'R'}) = E_o(a_I + a_{oI} + a_R)(a_{I'} + a_{oI'} + a_{R'}) \qquad (4.11)$$

Since the i and r facets are nested within o for the measurement procedure, taking an expected value over o automatically involves taking expected values over I, I', R, and R'. Therefore, the expected value over each of the crossproducts in Eq(4.11) is zero and the errors, D_{oIR}, are uncorrelated. Similarly, the covariance of the errors, d_{oIR}, for two administrations of the measurement procedure is also equal to zero.

In clasical test theory, errors of measurement are assumed to have an expected value of zero, to be uncorrelated across pairs of observations, and to be uncorrelated with the universe score. Errors of measurement that satisfy these requirements will be referred to as random errors. It is clear from the discussion just presented that a measurement procedure based on independent random samples from the universe of generalization satisfies these three requirements. Therefore, as long as an instance of the measurement procedure is defined as a randomly sampled observation from the universe of generalization, all of the effects contributing to the error of measurement for a dispositional attribute are random errors.

Cronbach et al.(1972) define a generalizability coefficient as the ratio of universe score variance, which is given in Eq(4.6), to the observed score variance, which is given in Eq(4.4).

$$Er^2(X_{oIR},u_o) = \frac{\sigma^2(o)}{\sigma^2_i(o) + \sigma^2(oi)/n_i + \sigma^2(i)/n_i + \sigma^2(r)/n_i n_r} \qquad (4.12)$$

where n_i is the number of conditions of the i facet sampled for each measurement, and n_r is the number of replications for each of these conditions.

The coefficient in Eq(4.12) incorporates tests of two separate invariance properties, one for the i facet and the other for the replication facet. Since the replication facet represents the effects of all but one of the facets in the universe of generalization, the second of the invariance properties is very general. If the observations in each objects of measurement are invariant over the i facet, the variance components for the i mean effect and the oi interaction must be small. Similarly, if observations are to be invariant over all other facets in the universe of generalization, the replication variance component must be small. In general, all of the variance components that appear as part of the error variance in generalizability coefficients are associated with invariance properties. To the extent that these variance components are close to zero, the invariance properties are good approximations.

Note that no assumptions need to be made about how the errors are distributed. In particular, it isn't necessary to assume a normal distribution for the errors in order to estimate generalizability coefficients. Assumptions about the distribution of errors are used in constructing confidence intervals, but confidence intervals will not be dicussed in this paper.

THE INFLUENCE OF SAMPLE SIZE ON GENERALIZABILITY

In classical test theory, the error variance is undifferentiated, and increasing the number of observations averaged to obtain an observed score leaves the true score variance unchanged and decreases the error variance. Where an observed score is defined as the mean over a sample of observations for the objects of measurement, increasing the size of this sample deceases the sampling variance of the mean. This regularity is the basis for the Spearman-Brown formula for changes in the length of a test.

For the coefficient in Eq(4.12), the relationship between the error variance and the number of conditions sampled for a facet is not so simple. However, by using Eq(4.12), it is possible to predict the generalizability coefficient for any number of conditions of the i facet and any number of replications. The fact that the generalizability coefficient can be predicted for various combinations of sample sizes for the different facets makes it possible to maximize the dependability of measurement for a fixed number of observations. In general, this is accomplished by sampling most thoroughly those facets that make the largest contribution to the error variance.

The fact that it facilitates the design of efficient measurement procedures is one important advantage of generalizability theory. However, an equally important point for the purposes of this paper is the fact that the total error variance for a measurement procedure can be made arbitrarily small by increasing the sample sizes. Therefore, if the variance component, $\sigma^2(o)$, is greater than zero, the generalizability coefficient in Eq(4.12) approaches a limit of 1.0, as the sample sizes for all facets approach infinity (For facets with a finite number, N_i, of conditions, the variance components for the facet go to zero as the sample size, n_i, approaches N_i). Therefore increasing the sample sizes for various facets provides a simple way of improving the dependability of any measurement procedure.

However, there are practical limits on how far this method can be pursued, and for important attributes, it is often impractical to achieve satisfactory dependability of measurement by increasing sample sizes. Later in this paper, more sophisticated approaches to the problem of measurement errors will be discussed. Although these techniques make it possible to control errors of measurement without inordinately large sample sizes, they also make it necessary to replace the simple universe sampling model discussed in this section by more complicated models.

A UNIVERSE SAMPLING MODEL FOR VALIDITY

Numerical estimates of generalizability coefficients are developed in two steps. First, components of variance are estimated in G studies. Second, generalizability coefficients are calculated using the estimated variance components and the sample sizes for the D study.

Since the universe score for each objects of measurement has been stipulated to be the value of the attribute for the objects of measurement, a measurement procedure is valid to the extent that it accurately estimates the universe score. For a measurement procedure consisting of random sampling from the universe of generalization, the observed score is an unbiased estimate of the universe score, and the random errors assumed in Eq(4.7) are the only sources of error in the measurement procedure. Since the generalizability coefficient provides an index of how accurately universe scores can be inferred from observed scores, it can be interpreted as a validity coefficient.

By definition, therefore, the value of the attribute for an object of measurement is the universe score, or the mean over all observations in the universe of generalization for the object of measurement. If this universe score were available, it would be a perfectly valid measure of the dispositional attribute. However, the universe score is generally not available and samples of observations must be used to estimate it. The primary requirement for measurement of an attribute is therefore that it provide accurate estimation of universe score for the attribute. This leads to the following definition of dispositional validity.

> A measurement procedure is said to be valid for a dispositional attribute to the extent that it provides dependable estimates of the universe score for the universe of generalization defining the attribute.

The validity of a measurement procedure is an index of the accuracy of inferences from a sample mean to the mean over the universe of generalization, where accuracy is defined by the expected squared error or by a coefficient of generalizability. Validity is a matter of degree, rather than an all-or-non property, and depends on both the measurement procedure and the attribute.

The value of the generalizability coefficient depends on how thoroughly the measurement procedure samples the universe of generalization, and this is determined by the design of the measurement procedure and by the definition of the attribute being measured. In particular, the more narrowly the universe of generalization is conceived, the more dependable the measurements will be. Cronbach, et al (1972, p.352) point out that, "investigators often choose procedures for evaluating the reliability that implicitly define a universe narrower than their substantive theory calls for. When they do so, they underestimate the 'error' of measurements, that is, the error of generalization".

In a sense, the definition of validity given above is similar to the classical notion of criterion validity, with the universe score being taken as the criterion. However, this similarity is relatively superficial. Unlike most of the criteria used with criterion validity, the universe score is an abstraction, a parameter defined on a universe of observations. Since the universe score is not directly observable, it isn't possible to estimate the validity by correlating observed scores with universe scores.

Although the universe score is an abstraction, and therefore not directly observable, it can be a relatively well-defined abstraction. To the extent that the universe of generalization is clearly defined, the accuracy achieved in estimating universe scores can be estimated by a generalizability coefficient, and therefore the validity of the measurement procedure can be represented by this coefficient. The clarity of definition of the universe of generalization is an issure that would be considered under the heading of content validity (Cronbach,1971).

Therefore, if the operational definition of a dispositional term were clearly specified, and if random samples could be drawn from the universe of generalization associated with this definition, validation would be relatively straightforward. Unfortunately the universe of generalization is usually not so clearly defined. This does not detract from the appropriateness of the definition of validity given above, but it does impose some limitations on the application of this definition. However these limitations are not new; they are closely related to the general problem of induction which arises in all scientific research. Establishing the validity of a measurement procedure requires the empirical verification of a number of invariance properties, and this task is not necessarily a simpler task than the verification of other empirical laws. The problem of induction that arises in verifying scientific laws and some of the solutions that have been proposed will be discussed more fully in a subsequent section.

The fact that a generalizability coefficient can be an index of validity may be surprising since generalizability theory is usually seen as an extension of reliability theory. However, the interpretation of Eq(4.12) as a validity coefficient is achieved only by making the strong sampling assumption that the observed scores are based on random samples from the universe of generalization (Tryon, 1957; McDonald, 1978). For most measurement procedures, observations are generalized to universes of generalization that are much broader than the universes from which they are sampled. It is not unusual, for example, for inferences to be drawn about broadly defined universes of behaviors on the basis of responses to a particular type of written test items. Similarly, a series of weight measurements may be obtained with a particular spring. In neither of these examples is it reasonable to assume that the observations are a random sample from the universe of generalization for the attribute being measured, and therefore a generalizability coefficient based on these observations would not be a validity coefficient.

Therefore, this simple universe sampling model presented in this section does not provide an adequate analysis of the validity of the great majority of

measurement procedures, which do not consist of random samples from their intended universe of generalization. For most of the attributes that are of interest in the behavioral sciences, standardization is used to control errors of measurement, which are often unacceptably large when observations are randomly sampled from the universe of generalization. Standardization involves an explicit decision not to use random samples from the universe of generalization in estimating universe scores. A standardized measurement procedure samples observations from a universe which is a subuniverse of the universe of generalization, and therefore requires a somewhat more sophisticated model for validity than that presented in this section.

Another method for controlling errors is the use of stratified sampling designs rather than simple random sampling. For example, in assessing the dependability of generalizations from the items on an achievement test to a universe of items, the assumption that items are randomly sampled within strata is undoubtedly much more realistic than the assumption of simple random sampling. (Stratified sampling is discussed by Rajaratnam, Cronbach, and Gleser, 1965.)

In general, an analysis of the dependability of a measurement procedure should reflect the sampling design for in the measurement procedure, and generalizability theory makes it possible to do this in a systematic way. Although the more realistic sampling models add complexity to generalizability analyses and may cause problems in estimation, the analysis of these more elaborate sampling designs is often especially informative; in a later section, convergent validity will be shown to be equivalent to a generalizability analysis with standardization of the method of observation.

Also, it is typically the case that there are unintended violations of the sampling assumptions in the G study. The effects of departures from the random sampling assumption cannot be estimated accurately, and therefore the interpretation of the results of G studies must always be somewhat tentative. The violation of sampling assumptions is, of course, a general problem in research, and the clouding of interpretations that results from such violations isn't unique to generalizability theory.

However, sampling problems are especially acute in the interpretation of generalizability coefficients because the estimation of these coefficients requires sampling from both a population of objects of measurements and a universe of generalization. Although the population, consisting as it does of the objects that are of primary interest to the researcher, is likely to be as well defined as other populations investigated in science, most universe of generalizations don't even meet this rather loose standard. The universe of generalization, which, by definition, consists of facets that are not being systematically investigated, is likely to be more poorly defined than the population.

Unintended violations of the sampling assumptions may introduce bias into the samples of some of the facets being investigated in a G study, and, given the universal applicability of Murphy's law, it would be unrealistic to assume that the estimates of variance components will be robust against such sampling biases. These considerations suggest, of course, that every effort should be made to avoid violations of the sampling assumptions. However, it would also seem prudent to include some explicit recognition of the possibility of sampling bias into the interpretations of generalizability coefficients. In the last section of this paper, it will be shown that if generalizability analyses are interpreted as tests of assumptions about invariance properties, it is possible to make these interpretations less vulnerable to violations of the sampling assumptions; the price to be paid for this increased security is a weakening of the conclusions drawn from various studies.

V Standardization and the Universe of Allowable Observations: One way to Refine the Sampling Model

As indicated earlier, the inclusion of an explicit theory of errors makes it possible for relatively simple theories to provide a consistent account of a wide range of observations. The inconsistency that would otherwise arise in these simple theories because of violations of invariance properties is accounted for by errors of measurement. Furthermore, the magnitude of these errors can be estimated, and the effects of such errors can therefore be taken into account in interpreting current observations and in predicting future observations.

Although the introduction of a theory of errors has several advantages, these advantages are most pronounced when the errors involved are small. The smaller the error variance, the more accurate the inferences that can be drawn from one observation to another or from an observation to the universe of generalization. It is desirable therefore that the error variance be as small as possible.

There are three ways to decrease the error variance and therefore to increase the precision of measurement. The first way to decrease the error variance is to base each measurement on a larger sample of observations from the universe of generalization. This approach is widely used in both the physical and behavioral sciences, and is discussed in detail by Cronbach et.al.(1972). One advantage of generalizability theory is that it indicates how to obtain the greatest increase in precision for a given increase in the number of observations sampled.

A second way to reduce errors is to restrict the universe of generalization. The more narrowly the universe of generalization is defined the smaller the errors will be. In the limiting case, if an observation is not generalized to any wider universe, but is interpreted as an observation, there is no error of measurement. Although narrowing the universe of generalization decreases the error variance, it can also limit the usefulness of the measurements, and is, therefore, not a panacea. This approach is discussed in the next section.

The third method for controlling errors of measurement is to standardize the measurement procedure. Standardization can be very effective in reducing errors of measurement, but it can also be misleading, and therefore requires careful examination. The remainder of this section is devoted to the implications of standardization.

STANDARDIZATION OF MEASUREMENT PROCEDURES

Since errors of measurement result from variations in the conditions of observation, these errors may be reduced by controlling, or standardizing, the conditions of observation. If the observations on an object of measurement vary as some facet varies, these observations may be made more consistent by making all observations with the same condition of the facet. If all applications of a measurement procedure employ a particular condition, or set of conditions, of a facet, the measurement procedure is said to be standardized on the facet.

Standardization of the i facet changes the design of the measurement procedure so that the objects of measurement are crossed with the same conditions, I^*, of the i facet for all measurements, but it doesn't alter the definition of the attribute. Standardization of a measurement procedure is not intended to imply a change in the universe of generalization, which continues to include the full universe of conditions for the i facet. Therefore, the universe score for object, o, is still u_o, as given by Eq(4.5).

The observed score for a measurement procedure with the i facet standardized can be represented by:

$$X_{oI^*R} = u + a_o + a_{I^*} + a_{oI^*} + a_R \qquad (5.1)$$

The expected value of the observed score over repeated application of the standardized measurement procedure is given by:

$$u_o^* = \underset{R}{E}(X_{oI^*R}) = u + a_o + a_{I^*} + a_{oI^*} \qquad (5.2)$$

For a standardized measurement procedure, therefore, the observed score is a biased estimate of the universe score, unless the last two terms in Eq(5.2) happen to be zero. This bias also appears in the errors for point estimates of universe scores, given by:

$$\begin{aligned} D_{oI^*R} &= X_{oI^*R} - u_o \\ &= a_{I^*} + a_{oI^*} + a_R \end{aligned} \qquad (5.3)$$

Eq(5.3) is the same as Eq(4.7) except for the fact that in Eq(5.3), the first term is a constant for all observations and the second term is a constant for all observations on a particular object of measurement; in Eq(4.7), all three terms are random variables. The expected value of the error, D_{oI^*R}, over repeated observations on the object, o, is given by:

$$\begin{aligned} \underset{R}{E}(D_{oI^*R}) &= \underset{R}{E}(a_{I^*} + a_{oI^*} + a_R) \\ &= a_{I^*} + a_{oI^*} \end{aligned} \qquad (5.4)$$

The amount of bias in the estimation of universe scores is indicated by the two terms on the right side of Eq(5.4).

The expected squared error for object, o, is given by

$$\underset{R}{E}(D^2_{oI^*R}) = (a_{I^*} + a_{oI^*})^2 + \sigma^2(R) \qquad (5.5)$$

Notice that the first term in Eq(5.5) involves the sum of two constants rather than a variance component, and that the second term is the variance component for replications.

The expected value of Eq(5.5) over the i facet is the same as the expected value of the squared error, D_{oIR}, for the unstandardized procedure, given by Eq(4.8). Therefore, standardization on a randomly chosen set of conditions of a facet does not decrease the expected squared error for point estimates of universe scores.

If I^* can be chosen so that $(a_{I^*} + a_{oI^*})^2$ is small compared to the sum of the first two variance components in Eq(4.8), the expected squared error for the standardized measurement procedure will be smaller than the expected squared error for the unstandardized procedure. A biased estimate with a small variance is often more useful than an unbiased estimate with a large variance. However, the problems of estimation involved in choosing a "good" value for I^* are substantial (See Cronbach et al, 1972, p. 101).

It may happen that there is no choice of I^* that significantly reduces the squared error, and if an I^* can be found with a small value for $a_{I^*}^2$, this choice may involve an unacceptably value for $a_{oI^*}^2$. Another possibility is to estimate the value of a_{I^*} by "calibrating" conditions of the i facet, and subtracting the estimated value of a_{I^*} from all observed scores; however, this is equivalent to using regression estimates with a slope of 1.0, and if this approach is to be used at all, it would probably be better to use standard regression estimates for the observed scores. The use of regression estimates of universe scores introduces a third type of error (Cronbach et al, 1972, p.106-107), which is not discussed in this paper.

Therefore, standardization may decrease the error for point estimates of universe scores, but does not necessarily do so. Standardization is a much more promising approach when observed scores are used to estimate universe scores relative to the average universe score in the population. If all observations have I^* as the conditions of the i facet, the expected value of the observed score over the population is:

$$u_{I^*} = u + a_{I^*}$$

and, therefore,

$$d_{oI^*R} = (X_{oI^*R} - u_{I^*}) - (u_o - u)$$

$$= a_{oI^*} + a_R \qquad (5.6)$$

The main effect, a_{I^*}, does not appear in Eq(5.6) because it is a constant for all observed scores, and therefore has no effect on the differences between observed scores and the mean observed score.

The expected value of d_{oI^*R}, over repeated applications of the standardized measurement procedure is given by:

$$E_R(d_{oI^*R}) = a_{oI^*} \qquad (5.7)$$

Therefore, the standardized measurement procedure is also biased in its estimates of universe deviations scores, but the magnitude of the bias, consisting only of the interaction effect, a_{oI^*}, is smaller than it is for point estimates of universe scores. Note that the expected value of this specific bias, over the population, is zero. Unless a_{oI^*} is zero for all objects of measurement, therefore, universe deviation scores are systematically overestimated for some objects of measurements and are systematically underestimated for others.

The expected value of the squared error over replications for object o is:

$$E_R(d_{oI^*R}^2) = (a_{oI^*})^2 + \sigma^2(R) \qquad (5.8)$$

The expected value of Eq(5.8) over all possible choices of I^* is:

$$E_{IR}E(d_{oI^*R}^2) = \sigma^2(oI) + \sigma^2(R) \qquad (5.9)$$

Therefore, the squared error, $d_{oI^*R}^2$, is expected to be smaller for the standardized measurement procedure than for the unstandardized measurement

procedure, given by Eq(4.10), even if standardization is on randomly chosen conditions of the i facet. Furthermore, if an I* is available for which the specific systematic error is particularly small, it is possible by standardizing the i facet to obtain a procedure with a small bias and with an expected squared error that is much smaller than that of the unstandardized measurement procedure. (Again, sampling problems make it very difficult to make an optimal choice for I*).

The main advantage in standardization is that it can be used to reduce error variance. In practice, standardization of a facet is most useful when observed scores are used to estimate universe deviation scores and the variance for the main effect of the i facet, $\sigma^2(I)$, is relatively large. Standardization of the i facet automatically eliminates $\sigma^2(I)$ from the error variance for comparative decisions, and if regression estimates are feasible, can eliminate $\sigma^2(I)$ from the error variance for point estimates. Although it is sometimes possible to choose conditions for the i facet so that the expected value of a^2_{oI*} over the population is small, this goal is not easy to achieve.

SYSTEMATIC ERRORS

Standardization is a powerful tool for reducing the magnitude of errors of measurement. As indicated above, however, the realization of benefits from this technique may require judicious selection of the conditions, I*, chosen for standardization, and this is not a trivial problem. Furthermore, there is a price to be paid for the benefits of standardization.

If the condition of the i facet is the same for all observations, the effects, a_{I*} and a_{oI*}, are constants over replications of the measurement procedure. For a given objects of measurement, therefore, standardization changes some components of the error of measurement from random variables to constants. Components of the error that are constant for all observations on an objects of measurement are called systematic errors. The effect, a_{I*} is a general systematic error since it a constant over all observations. The interaction effect, a_{oI*}, which is a constant for each objects of measurement but may vary from one objects of measurement to another, is a specific systematic error.

The systematic errors have neither of the two defining properties of random errors. First, the expected value of the systematic errors over repeated application of the standardized measurement procedure is not zero. Therefore, systematic errors introduce bias into estimates of the universe scores. The main effect for the i facet, is the same for all objects of measurement, and represents a general bias, which is present in D but not in d. The interaction effect, a_{oI*}, is a specific bias for each objects of measurement, o, and it affects both D and d. Since the systematic errors are constant for each objects of measurement, they do not tend to "cancel out" over a series of observations.

Second, the systematic errors are correlated across independent administrations of the measurement procedure. Since the expected value of Eq(5.4) over the population is a_{I*} the covariance between the errors, D, on two independent administrations of the standardized measurement procedure is given by: (Lord and Novick, p.181)

$$\mathrm{cov}(D_{oI*R}, D_{oI*R'}) = E_o(a_{oI*} + a_R)(a_{oI*} + a_{R'})$$

$$= \sigma^2(oI*) \qquad (5.10)$$

Thus, the errors of measurement for the standardized measurement procedure are correlated and the magnitude of the correlation depends on the magnitude of the specific systematic errors.

Similarly, the expected value of Eq(5.7) over the population is zero, and the covariance between the errors, d, on two independent administrations of the standardized measurement procedure is given by:

$$\text{cov}(d_{oI*R}, d_{oI*R'}) = E_o(a_{oI*} + a_R)(a_{oI*} + a_{R'})$$

$$= \sigma^2(oI*) \qquad (5.11)$$

Since the systematic errors are correlated across observations and do not have a mean of zero, they cannot be interpreted as the kind of random errors that appear in reliability coefficients. The interpretation of systematic errors raises issues usually associated with validity.

THE UNIVERSE OF ALLOWABLE OBSERVATIONS

In standardizing the i facet by requiring that every measurement involve the conditions $I*$, a new kind of universe, the universe of allowable observations, is introduced. The universe of allowable observations is a subset of the universe of generalization, and includes all obsevations in the universe of generalization that have the appropriate condition for each standardized facet. An instance of the standardized measurement procedure is a randomly sampled observation from the universe of allowable observations. By contrast, an instance of the unstandardized measurement procedure is an observation randomly sampled from the full universe of generalization. The universe of allowable observations defines a measurement procedure in the same way that the universe of generalization defines an attribute; both are "extensive" definitions.

Since standardization does not change the universe of generalization, measurements involve inferences to the universe of generalization rather than the universe of allowable observations. However, because it is easier to sample from the universe of allowable observations, an investigator who is evaluating a standardized measurement procedure will often begin by examining the dependability of inferences from observed scores to the mean over the universe of allowable observations. To do this, a G study can be conducted with observations randomly sampled from the universe of allowable observations.

If the i facet is standardized as $I*$ in the universe of allowable observations, all observations in the G study involve $I*$, and the i facet is a "hidden facet". The effects involving the i facet are confounded with other effects and cannot be estimated independently. An observed score can be written as

$$X_{oI*R} = (u + a_{I*}) + (a_o + a_{oI*}) + a_R \qquad (5.12)$$

where R is a replication index representing the combined effects of all facets other than the i facet. The terms enclosed in parenthesis in Eq(5.12) are completely confounded, and, in analyzing the G study, the model equation may be taken as:

$$X'_{oIR} = u' + a'_o + a'_R \qquad (5.13)$$

where

$$u' = u + a_{I*} \qquad (5.14)$$

$$a'_o = a_o + a_{oI*}$$

$$a'_R = a_R$$

Two variance components would be generated by a G study with the i facet fixed as I*, and these variance components can be written as:

$$\sigma^2(o') = \sigma^2(o) + \sigma^2(oI*) \qquad (5.15a)$$

$$\sigma^2(R') = \sigma^2(R) \qquad (5.15b)$$

where the variance components on the right side of Eq(5.15) are those that could be estimated in a G study in which conditions of the i facet are randomly sampled from the universe of generalization and are crossed with the objects of measurement.

The investigator who conducts a G study in which all observations are sampled from the universe of allowable observations will estimate the universe score variance by Eq(15a), and the expected observed score variance by:

$$\sigma^2(X) = \sigma^2(o') + \sigma^2(R)$$

$$= \sigma^2(o) + \sigma^2(oI*) + \sigma^2(R) \qquad (5.16)$$

Using Eqs(5.15a) and (5.16), the generalizability coefficient would be estimated as:

$$Er^2(X_{oI*R}, u^*_o) = \frac{\sigma^2(o')}{\sigma^2(o') + \sigma^2(R')}$$

$$= \frac{\sigma^2(o) + \sigma^2(oI*)}{\sigma^2(o) + \sigma^2(oI*) + \sigma^2(R)} \qquad (5.17)$$

Eq(5.17) assumes that generalization is over replications, and not over the i facet, and indicates the dependability of inferences from observed scores to the mean over the universe of allowable observations, u^*_o, in Eq(5.2). The dependability of inferences to the universe score is given by the ratio of the universe score variance in Eq(4.6) to the observed score variance for the standardized measurement procedure, given by Eq(5.16):

$$Er^2(X_{oI*R}, u_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oI*) + \sigma^2(R)} \qquad (5.18)$$

Eq(5.17) is approximately equal to the expected correlation, over the population, of two independent administrations of the standardized measurement procedure. Since Eq(5.17) indicates the consistency among the observed scores derived from the standardized measurement procedure, it can be interpreted as a reliability coefficient. Since Eq(5.18) reflects the agreement between observed scores and the value of the attribute as given by u_o, it can be interpreted as a validity coefficient.

RANDOM ERRORS AND RELIABILITY

The question of validity has been taken to be equivalent to the question of how well the results of a measurement procedure estimate the universe score. This question is answered by determining how well the results of the measurement procedure satisfy the invariance properties implicit in the operational definition of the dispositional attribute. Because standardization is so commonly employed in designing measurement procedures, however, the operational definition of the measurement procedure is usually not the same as the operational definition of the attribute; the observations that are actually used to estimate universe scores are typically drawn from a universe of allowable observations which is a sub-universe of the universe of generalization. A natural question to ask, then, is how well the results of particular instances of the measurement procedure generalize to the universe of allowable observations. This question is equivalent to the question of how well repeated administrations of the testing procedure, (i.e. repeated samples of observations from the universe of allowable observations) agree with each other. This issue is usually treated under the heading of reliability. Within the sampling model, reliability is defined in terms of the universe of allowable observations

> A measurement procedure is reliable to the extent that its observed scores provide dependable estimates of the mean over the universe of allowable observations.

Note that reliability is defined as a property of a measurement procedure, and does not depend on the definition of the attribute. As noted earlier, dispositional validity depends on both the measurement procedure and the attribute being measured. This distinction is consistent with the traditional definitions of reliability and validity. Reliability provides an index of consistency among the scores from independent administrations of a measurement procedure, and validity indicates the relationship between the results of a measurement procedure and an interpretation of these results which goes beyond the definition of the measurement procedure.

The reliability of a measurement procedure is an index of the consistency among the observed scores in the universe of allowable observations. This definition is equivalent to the definition of reliability for randomly parallel tests if the "true score" is equated with the mean over the universe of allowable observations. The reliability of a measurement procedure is limited by the magnitude of the random errors only; specific systematic errors tend to increase the reliability.

If the i facet is standardized, the interaction effect, a_{oI*}, is a systematic error, and is included in the numerator of the reliability coefficient in Eq(5.17). Since variance components are positive,

$$Er^2(X_{oI*R},u_o) \leq Er^2(X_{oI*R},u_o^*) \qquad (5.19)$$

$Er^2(X_{oI*R},u_o^*)$, which indicates the dependability of inferences from observations to the mean over the universe of allowable observations, it is a reliabilty coefficient, while $Er^2(X_{oI*R},u_o)$, which indicates the dependability of inferences from observations to the mean over the universe of generalization, is a validity coefficient. Therefore, the inequality in Eq(5.19) restates the well-known result from classical test theory that reliability is an upper bound for validity. For the sampling model, this result can be interpreted as reflecting the fact that generalization to the universe of allowable observations is always at least as dependable as generalization to the more broadly defined universe of generalization.

Although a G study in which the i facet is hidden does not address the central issue of the dependability of inferences from observed scores to universe scores, it is useful for two reasons. First, it provides an upper bound on the validity of the measurement procedure. A reasonably high value for the reliability coefficient, $Er^2(X_{oI*R}, u_o^*)$, doesn't establish the measurement procedure as having a high validity, but a low value can establish that the procedure has a low validity. In a subsequent section, the importance of such one-sided tests will be discussed. Second, the G study with the i facet fixed does provide an estimate of the random error variance, $\sigma^2(r)$, which is needed for estimating the validity coefficient, $Er^2(X_{oI*R}, u_o)$. If $\sigma^2(oi)$ is estimated in a subsequent G study, the validity coefficient could be estimated directly.

Of course, if the universe of generalization had only two facets, the investigator could estimate the variance components for both facets in a single study. The need for more than one G study arises from the fact that most universe of generalizations have many facets, and only a few facets can be systematically investigated in any G study. Since large sample sizes are generally necessary for the accurate estimation of variance components in designs with as few as two facets (see Smith, 1978), an adequate analysis of the generalizability of a measurement procedure will usually require a number of G studies.

SYSTEMATIC ERRORS AND VALIDITY

The difference between Eq(5.18), which has been interpreted as a validity coefficient, and Eq(5.17), which has been interpreted as a reliability coefficient, is in the role played by $\sigma^2(oI*)$. A reliability coefficient indicates the consistency of observed scores from one administration of a measurement procedure to another. As indicated by Eq(5.11), $\sigma^2(oI*)$ is the covariance of the errors of measurement over repeated observations on the object, o. Therefore, the covariance between the observed scores on two independent administrations of the standardized measurement procedure (two independent samples from the universe of allowable observations) increases as $\sigma^2(oI)$ increases. Therefore, as the magnitude of the specific systematic errors increases, the reliability increases. The magnitude of $\sigma^2(oI)$ provides information about how well the observations drawn from the universe of allowable observations correlate with the universe score for the attribute being measured. This would usually be interpreted as a question of validity.

In classical test theory, the "true score" for an objects of measurement is defined as the expected value of the observed score over repeated application of a measurement procedure to the objects of measurement. For a standardized measurement procedure, the expected value over repeated application of the measurement procedure implies taking an expected value over R, but not over I. Therefore, the mean over the universe of allowable observations is analagous to the true score of classical test theory. Although the standardized measurement procedure produces biased estimates of the mean over the universe of generalization, it does provide unbiased estimates of the mean over the universe of allowable observations. In the limit as the number of replications approaches infinity, the magnitude of the random errors approaches zero, and the observed score approaches u_o^*. Therefore u_o^* is a parameter for which the measurement procedure provides unbiased estimates. Since the measurement procedure is intended to provide estimates of the universe score, u_o, the correlation between u_o and u_o^* provides an index of the agreement between what the procedure actually estimates without bias and what it is intended to measure. The squared correlation between u_o^* and u_o is given by:

$$r^2(u_o^*,u_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oI^*)} \qquad (5.20)$$

Eq(5.20) represents the correlation between the universe score and an observed score for which the sampling of the universe of allowable observations is sufficiently thorough that random errors can be ignored. In addition, Eq(5.20) provides an upper bound on the validity of measurements based on a standardized measurement procedure. This correlation can also be obtained by taking the limit of Eq(5.18) as n_r approaches infinity, and therefore equals the limit of the squared correlation between the observed score and the universe score as the sample size for the observed score approaches infinity. For an observed score based on the average of a finite number of observations from the universe of allowable observations, the squared correlation between the observed score and the universe score will be less than or equal to Eq(5.20).

Eq(5.20) can be interpreted as a validity coefficient corrected for attenuation, and can be represented in a form analogous to classical attenuation formulas.

$$r^2(u_o^*,u_o) = \frac{Er^2(X_{oi*r},u_o)}{Er^2(X_{oI*R},u_o^*)} \qquad (5.21)$$

where the numerator of Eq(5.21) is the validity coefficient given by Eq(5.18), and the denominator is the reliability coefficient given by Eq(5.17). Therefore, $Er^2(u_o^*,u_o)$ represents a disattenuated validity coefficient for the standardized measurement procedure.

THE RELIABILITY-VALIDITY PARADOX

The inference from the observed score to the universe score can be decomposed into two parts. The first part is an inference from the observed score to the mean over the universe of allowable observations, and the second part is an inference from the mean over the universe of allowable observations to the mean over the universe of generalization, the universe score. The coefficient for inferences from observed scores to universe scores can be factored to represent the separate contributions of these two inferences:

$$Er^2_R(X_{oI*R},u_o) = Er^2_R(X_{oI*R},u_o^*)\ r^2(u_o^*,u_o)$$

$$= \frac{\sigma^2(o) + \sigma^2(oi)/n_i}{\sigma^2(o) + \sigma^2(oi)/n_i + \sigma^2(r)/n_i n_r} \cdot \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oi)/n_i} \qquad (5.22)$$

The first factor on the right side of Eq(5.22) is the reliability of the standardized measurement procedure and represents the dependability of inferences from observed scores to the expected value over the universe of allowable observations, u_o^*. The second factor on the right side of Eq(5.22) is a disattenuated validity coefficient and represents the dependability of inferences from u_o^* to u_o.

Assuming that the total number of observations, $n_i n_r$, is to be kept constant, the value of Eq(5.22) is maximized by maximizing n_i. The systematic error variance, $\sigma^2(oi)/n_i$, is inversely proportional to n_i, while the random error variance, $\sigma^2(R)$, is constant as long as the total

number of observations doesn't change. The validity of the standardized measurement procedure is improved by decreasing the sampling variance for the oi interaction effect, and this sampling variance is decreased by increasing the sample size for the i facet.

However, attending only to the reliability of the measurement procedure leads to the opposite conclusion. The reliability coefficient given by the first factor in Eq(5.22), is maximized by setting n_i equal to one, because this maximizes the oi interaction variance, $\sigma^2(oi)/n_i$, which is included in the numerator of the reliability coefficient. However this minimizes the disattenuated validity coefficient, which constitutes the second factor in Eq(5.22) and thereby minimizes the overall validity. Therefore, attempts to increase the reliability of the measurement procedure by standardizing a facet may decrease the validity. This phenomena has been called the reliability-validity paradox. A closely related phenomena called the attenuation paradox, is discussed by Loevinger(1954) and more recently by Lord and Novick(1968, p334).

CONVERGENT VALIDITY

Of the three main types of validity, construct, criterion, and content, dispositional validity is most similar to construct validity and may even be considered a part of construct validity. The definition of an attribute implies invariance properties. These invariance properties are laws that can be tested in G studies. If all of the invariance properties are tested and verified, the measurement procedure is valid. If some of the invariance properties are tested and no violations are detected, the validity of the procedure is partially supported. If even one invariance property is seriously violated then the procedure is invalid, and either the measurement procedure or the interpretation of the attribute must be revised.

The connection between invariance properties and construct validity can be made more explicit by examining how convergent validity (Campbell and Fiske, 1959) can be applied to dispositional attributes. Convergent validity of a measurement procedure can be investigated by letting the conditions of i represent different types of observations (objective tests, ratings, observation procedures, etc.). For each measurement a single condition from the i facet is sampled for all objects of measurement, and n_r replications are sampled independently for each object. The expected observed score variance for these measurements is

$$\sigma^2(X_{oiR}) = \sigma^2(o) + \sigma^2(oi) + \sigma^2(r)/n_r \qquad (5.23)$$

Since the universe score is defined as the expected value of the observed score over both the i facet and replications, the universe score variance is given by $\sigma^2(o)$. The generalizability coefficient for this measurement procedure is:

$$Er^2(X_{oiR},u_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oi) + \sigma^2(r)/n_r} \qquad (5.24)$$

The intraclass correlation coefficient in Eq(5.24) is approximately equal to the expected value, taken over the population of objects, and the universe of methods and replications, of the correlation between two sets of scores based on independently sampled methods. The interpretation of Eq(5.24) as a validity coefficient depends on the interpretation of the attribute; that is, it depends on the assumption that the attribute is not linked to a particular method. The value of Eq(5.24) can be used to check on the accuracy of a hypothesized invariance over methods.

Convergent validity is generally evaluated by measuring the same characteristic by several different methods, and estimating the correlations taken over objects, between the scores obtained on the various methods. If these correlations are high, then convergent validity is supported. Conversely, if these correlations are low, convergent validity is not suppported. The generalizability coefficient in Eq(5.24) is approximate equal to expected value of the correlation between observed scores obtained with pairs of methods randomly selected from the universe of generalization. It therefore provides a measure of the average convergent validity over all pairs of methods.

Assuming that $\underline{i}$ represents the method facet, the systematic errors a_{oi} are the object-method interactions. For a particular method $\underline{i}^*$, a_{oi^*} represents the specific systematic bias that results from using method I^*. Since the expected value of a_{oi^*} taken over all all conditions of the $\underline{i}$ facet is zero, $\sigma^2(oi)$ is the expected value of $a^2_{oI^*}$. A large value for $\sigma^2(oi)$ means that method has a serious effect on measurement. If $\sigma^2(oi)$ is zero, method has no influence on the measurement of deviation scores. Discriminant validity is discussed in a subsequent section.

VI Theory Development

In measuring an attribute, u_o, for the object, o, the effects, a_i and a_{oi} are components of the error. The larger these components are, the more difficult it is to obtain dependable measurements of u_o, and, therefore, the effects, a_i and a_{oi}, are generally viewed as nuisance factors to be reduced as much as possible. As described in the last section, standardization provides one way of dealing with these errors, but standarization introduces systematic errors. Furthermore, the fact that observations show a strong dependence on the i facet should be of interest in itself, aside from its effect on the inferences to u_o. Where this dependence can be described by an empirical law, a powerful technique for controlling errors of measurement becomes available.

The errors introduced into measurements of u_o by the i facet can always be eliminated by shifting attention to a new attribute, u_{oi}, which involves the same kind of operations that are used to define the attribute, u_o, but which has as its objects of measurement the pairs, oi, instead of the original objects of measurement, o. The universe scores, u_{oi}, for these new objects of measurements are found by taking the expected value of the observed scores over replications:

$$u_{oi} = \underset{r}{E}(X_{oiR}) = u + a_o + a_{oi} + a_i \qquad (6.0)$$

This redefinition of the objects of measurement changes a_i and a_{oi} from being part of the error to being part of the universe score. If the objects of measurements are defined by i as well as o, the difference between the two universe scores, u_{oi} and $u_{oi'}$, involving different conditions of the i effect, is taken as a substantive difference rather than as an error of measurement. Therefore, the two components, a_{oi} and a_i, which are equal to the difference between u_{oi} and $u_{oi'}$ become components of the universe score.

Measurements of u_{oi} are more dependable than measurements of u_o, because the interpretation of u_{oi} is narrower than that of u_o and thus involves inferences that are less susceptible to errors than those implied by u_o. While the original attribute, u_o, characterizes o for all conditions of the i facet, the new attribute, u_{oi} characterizes o for a particular condition of the i affect.

For each object of measurement, oi, the universe of generalization of the attribute, u_{oi}, includes observations with different conditions of the replication facet, but with constant values for o and i. Therefore, inferences from observed scores to u_{oi} involve generalization over r, but not over i. The universe of generalization for the attribute u_o includes observations with different values of both i and r, and inferences from X_{oiR} to u_o involves generalization over both the i effect and the r effect. (The generic term, "effect" is used in this section rather than the term "facet", because some of the objects of measurement being considered are defined by a combination of a condition of the o effect and the i effect. Therefore, neither the i effect nor the o effect separately specifies the objects of measurement. However, it is also true that the i and o effects are not facets of the universe of generalization.)

If the observed score, X_{oiR}, is used to estimate u_{oi}, the expected value over replications, the only source of error will be the replication facet. Therefore, the dependability of inferences from X_{oiR} to u_{oi} is given by:

$$Er^2(X_{oiR},u_{oi}) = \frac{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i)}{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i) + \sigma^2(R)} \qquad (6.1)$$

The coefficient in Eq(6.1) is approximately equal to the expected value of the squared correlation between the observed score, X_{oiR}, and the universe score, u_{oi}.

When the universe of generalization is restricted to a particular condition of the i effect, u_{oi} becomes the universe score, and Eq(6.1), which reflects the dependability of inferences from X_{oiR} to u_{oi}, is a validity coefficient, with the oi combinations as the objects of measurement and generalization over R. This validity coefficient $Er^2(X_{oiR},u_{oi})$ is never less than the validity coefficient, $Er^2(X_{oiR},u_o)$, with conditions of the o effect as the objects of measurement, and generalization over i and R. Restricting the universe of generalization improves the validity of measurement whenever $\sigma^2(oi)$ or $\sigma^2(i)$ is greater than zero. (By contrast, for nonzero values of $\sigma^2(i)$, standardization of the i facet always improves reliability but does not necessarily improve validity.)

The increase in validity obtained by restricting the universe of generalization is part of a trade-off in which decreases in the errors of measurement are obtained by narrowing the interpretation that can be given to the attribute. A high value for Eq(6.1) indicates that an inference from the observed score, X_{oiR}, to the universe score, u_{oi}, is dependable, and therefore provides justification for such inferences. If Eq(6.1) is to be taken as a validity coefficient, generalization must not go beyond the restricted universe of generalization, which involves a particular value of i; that is, inferences are to the universe score, u_{oi}, for the restricted universe of generalization in which both o and i are constants for all observations. The value of Eq(6.1) does not indicate the dependability of inferences from an observed score obtained with one value of i to universe scores involving different values of i. In particular, a high value for $Er^2(X_{oiR},u_{oi})$ doesn't justify inferences from u_{oi} to $u_{oi'}$, the universe score for the same value of o and a different value of i, or to u_o, the expected value of u_{oi} over all values of i. Therefore, a high value for $Er^2(X_{oiR},u_{oi})$ provides support for only a relatively limited set of inferences.

The expected correlation between u_{oi} and $u_{oi'}$, where i and i' are independently sampled conditions for each observation, is given by:

$$Er(u_{oi},u_{oi'}) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i)} \qquad (6.2)$$

Eq(6.2) is also equal to the the expected squared corelation $Er^2(u_{oi},u_o)$, between the universe score, u_{oi}, for universes that are restricted to a specific condition of the i effect, and the universe score, u_o for a universe that includes the i effect as a facet.

If the universe scores, u_{oi}, do not vary much as a function of i, then $\sigma^2(oi)$ and $\sigma^2(i)$ would be small, and the coefficient in Eq(6.2) would be close to 1.0, indicating that u_{oi} is a dependable estimate of $u_{oi'}$. Therefore fixing the value of the i facet isn't a serious limitation if the observed scores don't vary much over the i facet. However, this invariance of u_{oi} with respect to i would also imply that the validity coefficient in Eq(6.1) would not be substantially larger than the validity coefficient for the original universe of generalization, given by Eq(4.12), and there would be little advantage in restricting the universe of generalization.

Eq(6.2) can also be derived by setting n_i equal to one in Eq(4.12), and taking the limit as n_r approaches infinity,

$$Er^2(u_{oi},u_o) = \lim_{n_r \to \infty} Er^2(X_{oiR},u_o)$$

That is, Eq(6.2) provides an index of dependability of inferences to u_o for observed scores based on a single condition of the i facet and an infinite number of replications. Furthermore, by comparing Eqs(4.12), (6.1), and (6.2), it is clear that:

$$Er^2(X_{oiR},u_o) = Er^2(X_{oiR},u_{oi})\ Er^2(u_{oi},u_o) \qquad (6.2a)$$

Eq(6.2a) partitions the generalizability of inferences from X_{oiR} to u_o (or to u_{oi}) into two parts. The first part, $Er^2(X_{oiR},u_{oi})$, represents the dependability of inferences from X_{oiR} to u_{oi}, the mean over replications for a fixed value of i. The second part, $Er^2(u_{oi},u_o)$, is the dependability of inferences from u_{oi} to u_o. For the investigator who intends to generalize to the universe score, u_o, therefore, there is no benefit in fixing the condition of the i facet. As a matter of fact, the dependability of inferences to u_o would be improved by explicitly recognizing the i effect as a facet, and increasing n_i.

The main benefit derived from restricting the universe of generalization is the increase in the validity of measurement, the dependability of inferences from observed scores to universe scores. The main disadvantage associated with restricting the universe of generalization, is that it can lead to a very large increase in the number of objects of measurements in the population. If there were N_o objects in the original population, which can each be paired with N_i conditions of the i facet, there are N_oN_i objects in the new population. Unless N_i is very small therefore, the number of universe scores that need to be estimated for a complete description of the population may be greatly increased by restricting the universe of generalization.

INFERENCES THAT GO BEYOND THE SAMPLING MODEL

If the dependence of observations on the conditions of the i facet involved in the observations is investigated, it may be possible to characterize the conditions of the i facet by an attribute, w_i, such that:

$$u_{oi} = f(v_o,w_i) \qquad (6.3)$$

where f represents some function. Eq(6.3) expresses u_{oi} as a function of two variables. The component, v_o, depends on the value of o but does not depend on the value of i. The component, w_i, depends on the value of i but does not depend on the value of o. These two variables may be previously defined attributes or they may be defined as derived attributes by the law in Eq(6.3). Since the basic reason for considering the attribute u_{oi} in place of the attribute, u_o, is the lack of dependability in estimates of u_o, the universe score, u_o, is not a very promising candidate for the variable, v_o; if the estimates of v_o contain large errors of measurements, it may be impossible to identify an appropriate functional form, f, for Eq(6.3).

It is often the case therefore, that u_{oi} is expressed in a law with the following form:

$$u_{oi} = g(u_{oi*},\ w_i - w_{i*}) \qquad (6.4)$$

where g represents some function, and i* is a particular condition of the i facet. A new variable, u_{oi*}, is defined by restricting the universe of generalization for all objects of measurements to the fixed reference condition, i*, for the i effect. This new variable can be substituted for v_o because it is a function of o but not of i. Measurements of u_{oi} and of u_{oi*} are more dependable than measurements of u_o because they do not assume invariance over the i effect; this facilitates the development laws of the form given by Eq(6.4).

If a law like that in Eq(6.4) can be developed, the limitation inherent in measurements of u_{oi} can be overcome. With the help of Eq(6.4), measurements of u_{oi*} provide information about all conditions of the i effect for which w_i is known. This information is provided by the empirical law in Eq(6.4). Since the use of Eq(6.4) requires measurement of the attribute w_i for all values of i, this kind of inference from observations involving one condition of the i effect to what would be expected for observations for another condition of the i effect is more difficult to develop than inferences that use only invariance properties. However, this more complicated approach provides a detailed analysis of the relationship between o and i. Useful as they are, invariance properties do not provide such an analysis.

The empirical law, given by Eq(2.2), relating length to temperature is an instance of the kind of law indicated by Eq(6.4). This law can be used to improve the dependability of the inferences involved in measurements of length. A new quantity, l_{rt}, may be defined as an attribute of a rod at the temperature, t:

$$l_{rt} = L(rt) \qquad (6.5)$$

The object of measurement for l_{rt} is a rod-temperature combination, rt, instead of a rod. All observations that are used to estimate l_{rt} must involve a specific rod and a specific temperature, while the observations used to estimate l_r must involve the rod defining the object of measurement, but may involve a variety of temperatures. The attribute, l_{rt} has a smaller universe of generalization than the attribute l_r, and the direct interpretation of measurements of l_{rt} are restricted to the temperature specified in rt.

This restriction is effectively eliminated by the law of thermal expansion. For a set of rods that are made of the same material and for a fairly wide range of temperature, the coefficient of thermal expansion is a constant, k, and Eq(2.2) can be written as

$$l_{rt} = l_{rt*} + k(t - t^*)l_{rt*} \qquad (6.6)$$

where t* is some fixed reference temperature (For convenience, t* is often taken to be 20°C, a comfortable value for room temperature). Because temperature variations introduce error into measurements of l_r, (i.e. $\sigma^2(rt)$ is not zero), l_{rt*} can be measured more dependably than l_r. Also, the temperature differences, (t-t*), can be measured very accurately, and Eq(6.6) provides a very good fit to data over a wide range of temperatures. Therefore, the dependability of estimates of l_{rt} based on Eq(6.6) is limited mainly by the dependability in estimates of l_{rt*}. Therefore, fixing the temperature for measurements of length does not seriously limit the interpretation of these measurements.

The observations involved in measurement are of interest mainly because they support inferences to other observations. These inferences are of two kinds. First, there is an inference from the observation to the universe score, the mean over the universe of generalization. Second, there are inferences from one universe score to other universe scores.

Using the bridge analogy of Cornfield and Tukey(1956, p.912), these two inferences can be represented by the two spans of a bridge that crosses a river. The first span represents inferences from the observed score to a universe score, and the second span represents inferences from the universe score to the universe scores for other attributes.

If a well articulated theory is available connecting an attribute to other attributes, the second span, which is supported by empirical laws, may be made quite long without weakening the total inference. Laws of the kind indicated by Eq(6.4) make it profitable to shorten the length of the first span by narrowing the universe of generalization. Inferences from observed scores to the universe score, u_{oi*}, for the restricted universe of generalization have a higher validity than inferences to u_o. Therefore, restricting the universe strengthens the first span. A well confirmed law of the type given in Eq(6.4) provides a strong second span by justifying inferences from u_{oi*} to the other universe scores u_{oj}. Therefore restricting the universe of generalization does not result in any loss in generality; the second span is simply bearing a larger share of responsibility for the inferences based on observed scores. "A proposal to sample items from a broad domain at random is generally but not always a sign that one's understanding is crude" (Cronbach et al, 1972)

Note that an invariance property is a special case of the class of laws indicated by Eq(6.4). In particular, if the function, g, is such that u_{oi} is a constant for all conditions of the i facet (i.e., w_i is a constant), then u_{oj} is invariant with respect to the i facet. In such cases, there is no loss involved in taking o, instead of oi, as the object of measurement and there is some gain in simplicity. (In practice, it is often convenient to assume that u_{oj} is invariant with respect to the i facet, even where this assumption is known not to hold exactly).

A NOTE ON LATENT TRAIT MODELS

In the behavioral sciences, when a test consisting of some set of items is administered to a person, the observed score is usually interpreted as an estimate of the universe score for a disposion, with generalization over the item facet. Latent trait theory can be considered a special case of the kind of model being discussed here. For example, the Rasch model (Wright and Douglas, 1977) represents the probability, X_{pi}, that person, p, answers item, i, correctly in terms of an attribute, v_p, of the person and an attribute, w_i, of the item.

$$X_{pi} = \frac{e^{(v_p - w_i)}}{e^{(v_p - w_i)} + 1} \qquad (6.7)$$

The ability parameter is assumed to be an attribute of the person, and may vary from one person to another. However it is explicitly assumed that v_p does not vary with the sample of items used to estimate v_p. Furthermore it is at least implicitly assumed that v_p does not change as other conditions of observation vary. Similarly, items are the objects of measurements for the difficulty attribute, w_i, and the value of w_i is assumed to be independent of the sample of persons used to estimate w_i, and of the conditions of observation that may hold when w_i is estimated.

Latent trait theories do not generally incorporate an explicit theory of errors. The kind of inconsistencies that can be attributed to errors of measurement, are interpreted in terms of a lack-of-fit for latent trait models.

NOMOLOGICAL NETWORKS AND THE IMPORT OF MEASUREMENTS

The development of inferences beyond the universe of generalization for an attribute, as exemplified by the use of Eq(6.4), raises the question of the role of empirical laws in evaluating measurement procedures.

In discussing this issue, it is useful to define a third property of measurement, in addition to reliability and validity. This third property, the theoretical import, or the import, of measurement, can be defined as the total significance of what can be inferred from the measurement (Hempel, 1952). Import emphasizes the scope or range of inferences that can be drawn from a measurement as well as the accuracy of individual inferences. As defined here, import does not involve a numerical index, and it is not assumed that the import of individual inferences can be measured on any scale of utilities.

Hempel (1952,p.46) provides a good example of the distinction between "theoretical import" and and "empirical import"(or validity):

> Concepts with empirical import can be readily defined in any number, but most of them will be of no use for systematic purposes. Thus, we might define the hage of a person as the product of his height in millimeters and his age in years. This definition is operationally adequate and the term "hage" thus introduced would have relatively high precision and uniformity of usage; but it lacks theoretical import, for we have no general laws connecting the hage of a person with other characteristics

Although, "hage" lacks import, it is possible to measure "hage" with a high degree of reliability and validity.

The invariance properties, which provide the basic justification for interpreting observations as measurements, provide a core of import to all measurements. These invariance properties justify inferences from the observed scores to the universe score, and also, to some extent, to all other observed scores in the universe of generalization. If the attribute isn't involved in any other empirical laws, these inferences to the universe of generalization define the total import of the measurement.

The contribution of the invariance laws to import depends on the generality of these laws. For example, the import provided by invariance properties for the measurement of u_{oi} may be relatively minor, since inferences from X_{oiR} to scores for other values of i is not justified by these invariance properties. If attention is restricted to the invariance properties implied by the definitions of their respective universe of generalizations, the attribute, u_o has greater import than the attribute, u_{oi}. As such, measurements of u_{oi} do not justify inferences to other conditions of the i facet. Measurement of u_o, on the other hand, involves generalization over the i facet.

Measurable attributes that play a central role in the fundamental theories of a science are seen as having greater significance, or import, than attributes which are involved in one or two isolated empirical laws, or in no laws at all. The extended network of laws, which specifies the empirical content of the theory and also provides confirmation for the theory, can greatly extend the implications of measurements.

In practice, the development of empirical laws can lead to simultaneous increases in both the validity and the import of measurements. This is done by partitioning the universe of generalization into a number of more narrowly defined subuniverses, while connecting the universe scores for these

subuniverses through empirical laws. Physics has used this strategy very effectively. The history of the measurement of such basic measurable attributes as length reveals the gradual refinement of their universes of generalization. In the course of this development many facets have been standardized, including the physical object defining the unit of length, the temperature, and numerous aspects of procedure. At the same time, the import of length measurements has been increased by the use of nomological networks including Euclidian and non-Euclidian geometries, classical mechanics, and the theory of relativity. Using these theories, inferences can be drawn at one extreme, about the distances between galaxies, and, at the other extreme, about the sizes of subatonic particles, while defining length in terms of operations involving a particular platinum-iridium bar in Paris.

NOMOLOGICAL NETWORKS AND VALIDITY

It is clear therefore that the existence of empirical laws, and more importantly, nomological networks, may greatly extend the range of inferences that can be drawn from measurements. Therefore such networks can greatly increase the import of measurements. But returning to the question posed earlier, what are the implications of such networks of laws for the validity of measurement? In particular, what are the implications of an empirical law, like Eq(6.6), for the validity of measurement?

The definition of dispositional validity states that a measurement procedure is valid to the extent that its observed scores provide dependable estimates of universe scores. If this definition of validity is accepted, the existence of empirical laws relating measurements of an attribute to measurement of other attributes has no direct bearing on the validity of measurements of the attribute. Campbell (1921, pp.109-134) distinguishes clearly between the development of an acceptable measurement procedure and the application of this MP in the discovery of empirical laws. Campbell (1921, p.134) recognizes that it is, "because true measurement is essential to the discovery of laws that it is of such vital importance to science", but he does not use these laws to justify particular MPs. Measurement procedures are justified by a careful examination of the operations involved in the MP and by experimental verification of invariance properities.

The dependability of estimates of universe scores depends on the definition of the universe of generalization, the magnitude of variance components, and the extent of sampling of various facets for each observed score. For example, assuming that $\sigma^2(oi)$ is greater than zero, generalization from X_{oiR} to u_{oi} will be more dependable than generalization from the same observed score, X_{oiR}, to the more broadly defined universe score, u_o. However the dependability of estimates of u_o can be improved simply by sampling the i facet more thoroughly. In general, measurements of u_o can always be made as valid as measurements of u_{oi}, by increasing sample sizes.

All derived attributes and most basic attributes are involved in empirical laws, and these laws may be tied to the body of laws and theoretical constructs associated with a theory. For example, the law of thermal expansion, Eq(6.6), describes phenomena that physical theory would be expected to explain, and at least a partial explanation of these phenomena can be provided in terms of the motion of molecules. However, the existence of such an explanation is not necessary in order to interpret a coefficient of thermal expansion. With or without a theory, a coefficient of thermal expansion is interpreted as a measure of the degree to which a rod will expand when heated and contract when cooled. A model for the molecular structure of solids may provide insight into why this phenomenon occurs, but such models are not necessary for an understanding of what the phenomenon is.

However, networks of empirical laws do have a great, but indirect, influence on the validity of measurements. The laws in the network make it feasible to restrict the universe of generalization for attributes and therefore to increase the validity of measurements without decreasing their import. These more narrowly defined attributes depend on the network for their import, rather than on the invariance properties, and the magnitude of the errors of measurement are reduced because assumptions are made for fewer invariance properties.

THE TRADEOFF BETWEEN VALIDITY AND IMPORT

If the issue of import is ignored, it is easy to develop attributes that can be measured with a high degree of validly. It is only necessary to define the universe of generalization very narrowly. For narrowly defined universes, the inferences from observations to the universe involves generalization over relatively few facets, and in such cases estimates of the universe scores are likely to be very dependable. In the extreme case, where observations are interpreted simply as observations, there is no inference, and estimates of the universe scores are perfectly accurate.

However, researchers cannot ignore the issue of import, and decisions which involve tradeoffs between the validity of measurements and the import of measurements must be made. The researcher who interprets observations narrowly will draw more accurate inferences than the researcher who interprets observations broadly, but the inferences of the first researcher say less about the world than the inferences of the second researcher. The choice between narrow but dependable interpretations and broader but less dependable interpretations is a choice of strategy. The continuum of available strategies is more-or-less anchored by strict operationism (Bechtolt, 1959) at on end and by construct validity (Cronbach and Meehl, 1955) at the other end.

Strict operationism demands that attributes be defined narrowly enough to insure that the validity of the interpretations is essentially perfect. The strict operationist is unwilling to give any hostages to the future in the form of invariance properties that might turn out to be only approximations. If taken seriously, this position implies that observations should not be assumed to be generalizable to any wider universe of generalization, but should instead be interpreted simply as observations. All of the implications of the observation are to be derived by developing empirical laws that state relationships between observations. Strict operationalism is the strategy of pure empiricism, and theory plays essentially no role.(Bechtoldt, 1959)

Construct validity, in its most general form, would define an attribute in terms of all of the relationships in which the attribute appears. From the standpoint of construct validity, the definition of an attribute entails certain laws, and, in order for a MP to be valid its observed scores must satisfy these laws. These assumptions are the postulates of a theory, and may state relationships between the construct and other constructs, in addition to some set of invariance properties.

CONSTRUCT VALIDITY

It was stated earlier that reliability is a property of a measurement procedure, while validity is a property of a measurement procedure and the attribute being measured. Construct validity is defined as a property of a measurement procedure, a construct, and the network defining the construct.

According to Cronbach and Meehl(1955, p. 257):

> "Acceptance," which was critical in criterion-oriented and content validities, has now appeared in construct validity. Unless substantially the same nomological net is accepted by the several users of the construct, public validation is impossible. ... A consumer of the test who rejects the author's theory cannot accept the author's validation. He must validate the test for himself, if he wishes to show that it represents the construct as he defines it.

Therefore, in construct validity, it would not be completely accurate to say that a measurement procedure is valid for a construct. Instead, a claim for construct validity should specify a measurement procedure, a construct, and the theory defining the construct.

If constructs are defined in terms of a network, changing the network of laws in which the construct is embedded implies a change in the definition of the construct. Therefore, evidence for the construct validity of a measurement procedure may not apply in different networks. The invariance properties that are tested in validating a dispositional attribute form a subset of laws that may be part of many theories. Evidence for dispositional validity of a measurement procedure applies in all networks that involve the same definition of the attribute and therefore the same set of invariance properties.

As stated earlier, all of the procedures included in dispositional validity are consistent with construct validity. Indeed, the procedures suggested here form a subset of those proposed by Cronbach and Meehl(1955). The difference between the two approaches is in the fact that, within construct validity, the testing of any law involving a construct could be construed as a test of the validity of a measurement procedure for the construct. Dispositional validity restricts its attention to invariance properties; the testing of laws other than the invariance properties is seen as being directly relevant to an attribute's import but not to its validity.

None of the many types of research included within construct validity is excluded by the definitions in this paper. The major change being proposed is in how some studies will be interpreted, rather than in the kinds of studies to be done. In particular, the whole spectrum of research that could be interpreted in terms of the import of an attribute, would, within construct validity, be interpreted in terms of the validity of a measurement procedure.

Therefore, the apparent loss involved in giving up evidence from some parts of a nomological network is not really a loss at all. For dispositions, the inferences that are subsumed under construct validity are divided into two categories, validity and import. None of the studies encouraged by construct validity is thrown away in considering dispositions, but the results of some of these studies are interpreted as evidence for the import, rather than the validity, of the disposition. Although the evidence that can be used to support dispositional validity is more restricted than the evidence permitted by construct validity, the inferences that are drawn by dispositional validity are more restricted than those drawn by construct validity.

Construct validity is a property of measurement procedure in relation to a network because a measurement procedure is said to have construct validity by virtue of its inclusion in a validated network. For dispositional validity, import is a property of the network as a whole, and validity is a property of a measurement procedure in relation to the universe of generalization defining a disposition.

DISCRIMINANT VALIDITY

Discriminant validity can be examined by estimating variance components. A large variance component for the interaction between attributes and objects of measurement, would indicate that the intercorrelations among different attributes are small. This implies that the universe score for one attribute doesn't provide a dependable estimate of the universe scores for other attributes, and therefore that each of the attributes being measured provides information which is independent of the information in the other attributes. However, as applied to dispositional attributes, discriminant validity is more closely associated with the import than with validity.

The logic of discriminant validity depends on the existence of, at least, a rudimentary theory. If two attributes, A_1 and A_2, represent hypothesized constructs that are assumed by some theory to be unrelated, it would be expected that measurements of these attributes should also be unrelated. By the logic of construct validity, a strong relation between measurements of A_1 and A_2 could be interpreted as evidence that at least one of these two sets of measurements is invalid.

This kind of logic is not generally appropriate for dispositional attributes, because dispositions are defined in terms of universes of observations, and do not depend on theoretical networks for their meaning. Assuming that A_1 and A_2 are clearly defined dispositions, a strong relationship between measurements of these two attributes would be interpreted as an empirical law that theory would be expected to explain. In particular, a high correlation between measurements of A_1 and A_2 would be evidence against any theory which treated these two dispositions as being unrelated. It would not be interpreted as evidence for a lack of validity in the measurement procedures.

The exact interpretation of the relationship between A_1 and A_2 would, of course, depend on how the two attributes are defined. If the two universes had a high degree of overlap in their observations, it would not be surprising that the means over these two universes are related. However, even if there were no overlap in their universes and no other reason to expect a relationship, the two attributes could be strongly related without having the validity of their measurement procedures denied.

For example, the empirical result that there is a strong correlation between the thermal conductivity and electrical conductivity is not taken as evidence against the validity of the procedures used to measure these two dispositional attributes. Instead, the relationship between these two attributes is a legitimate empirical finding, which a theory of solids would be expected to accomodate.

It is generally true, however, that a very high correlation between different attributes may limit their usefulness for various purposes. If two attributes were perfectly correlated, the measurement of either attribute could serve all of the purposes served by measurements of both attributes. Therefore, the import of a measurement procedure depends on its having discriminant validity.

VII An Overview of Sampling Models for Validity

The earlier sections of this paper have discussed a sampling model for dispositional attributes in some detail. This section provides a general summary of this sampling model, and briefly discusses some issues, including objections to sampling models, not covered in the previous sections.

A GENERAL SUMMARY OF THE SAMPLING MODEL

The sampling model for the validity of measurements of dispositional attributes is based on generalizability theory. The "true" value of a dispositional attribute is the universe score, defined as the mean over the universe of generalization. A measurement procedure is valid for a dispositional attribute to the extent that the observed scores generated by the procedure are dependable estimates of the universe score.

The basic premise of sampling models is that measurement involves inferences from observed scores, which are based on samples of observations, to the mean over the universe from which these samples are drawn; for these inferences to be justified, a set of invariance properties must hold, at least approximately. The invariance properties must be verified empirically.

Dispositional validity could be estimated directly by using samples of observations, randomly sampled from the universe of generalization, to estimate a generalizability coefficient, which reflects the dependability of inferences from observed scores to universe scores. Although this direct approach has the virtue of simplicity, it is not practical in most cases.

The sampling model is based on a small number of assumptions. The sampling model makes no assumptions about underlying constructs; it neither affirms nor denies the existence of such theoretical constructs. The model makes no asssumptions about the distribution of observed scores, the distribution of universe scores, or the relationships between different kinds of observed scores. No restrictions are put on the universe of generalization or the universe of allowable observations except that the universe of allowable observations is, by definition, included within the universe of generalization; in particular, the model doesn't dictate what kinds of conditions can be defined as facets. The one assumption that is necessary for the sampling model is the random sampling assumption.

OBJECTIONS TO THE SAMPLING MODEL

The simple version of the sampling model assumes that an attribute is defined in terms of a universe of generalization and that measurements are based on random samples from this universe. This simple version of the sampling model ignores the complex sampling designs that are actually employed by measurement procedures. In practice, the assumption that observed scores are based on random samples from the universe of generalization does not apply whenever any facet is explicitly or implicitly standardized.

A number of authors (Loevinger, 1965; Rozeboom, 1966; Gillmore, 1979) have objected to the simple version of the sampling model because behavioral measurements do not generally consist of random samples from a clearly defined universe of generalization. Ambiguity in the definition of the universe of generalization is not unique to the social sciences. The discussion of attributes in an earlier section of this report made a point of emphasizing many of the sampling problems associated with measurement in the physical sciences. For most attributes, the boundaries of the universe of generalization tend to be quite fuzzy, and the sampling of conditions of various facets is far from random. It is generally impossible to select random samples from the universe of generalization, in part, because of vagueness in the definition of the universe of generalization, and, in part, because of practical difficulties.

All of these objections emphasize the problems inherent in trying to take random samples from the universe of generalization. In order for statistical inferences from observed scores to universe scores to be unbiased, the sample of observations must be selected randomly from the universe of generalization, and the simple version of the sampling model assumes that observations are randomly sampled from the universe of generalization.. To the extent that the random sampling assumption is untenable, the simple version of the sampling model is untenable. Since it is usually not possible to sample the universe of generalization randomly, this simple model is seldom applicable.

The sampling model developed in this paper recognizes that the estimation of a universe may involve a relatively complicated set of inferences. Within this more realistic model, inferences from observed scores to universe scores may be analyzed into several steps and may require the confirmation of an extensive network of laws for their justification. In particular, the reliability and validity of a measurement procedure may involve a large number of invariance properties. The investigation of import would add the other empirical laws to the network.

In light of the substantial difficulties in drawing random samples from the universe of generalization (not to mention the demands on sample sizes for the estimation of variance components, when random sampling is possible), how is this network to be verified? In the physical sciences, the verification of the required set of empirical laws is not generally accomplished in a single study. It is even less likely that behavioral measurements will be thoroughly evaluated in a single study, in which observations are randomly sampled from the universe of generalization.

Even a cursory review of the issues involved in the confirmation of laws would go far beyond the scope of this paper. However a few general remarks on this topic may put the problems involved in testing the invariance properties into perspective.

Karl Popper(1965,1968) has suggested an approach to the verification of laws, which accurately reflects the practice of science, and has been widely accepted. Popper views laws as conjectures which can be tested in various ways, but which can never be definitely confirmed. A general law can be applied to a large number of observations, and any of these observations can be used as tests of the law. A deterministic law that fails a single test or a statistical law that fails a large proportion of its tests is refuted. A law which is subjected to large number of tests of various kinds without being refuted is considered to be supported by these tests.

There is a clear lack of symmetry in the treatment of laws as conjectures that are subject to refutation; a law can be refuted in a single study, but, in general, even a large number of studies cannot definitely confirm the law. The more tests of various kinds that a law has been exposed to without being refuted, the more strongly it is considered to be supported, but the law is never completely confirmed. In a sense, therefore, Popper replaces the concept of the confirmation of a law by the concept of the degree of confidence in the law. Each successful empirical test of the law increases confidence in the law, and the failure of one test can cause the law to be refuted.

Toulmin (1953) describes physical laws not as inductive generalizations but as rules of inference, which can be used to draw conclusions from observed facts. The question to be asked about such rules of inference is not whether they are true or not, but how widely do they apply. Toulmin analyzes the role of laws somewhat differently from Popper, but, for the purposes of this paper, these two views are complementary. The presumptions that are made about the

class of observations to which a law applies are tested every time the law is applied, and therefore such presumptions are conjectures which are subject to refutation.

Both of these analyses of the methodology for the confirmation of scientific laws have implications for the validiation of measurement procedures. The definition of an attribute involves a universe of generalization. The claim that a measurement procedure generates valid measurements of the attribute is equivalent to the conjecture that the observed scores are invariant with respect to random sampling from the universe of generalization. This conjecture can be decomposed into a number of more specific invariance properties, each of which applies to a specific facet.

A successful test of any one of the invariance properties, which is based on a random sample of conditions from a facet, provides strong support for the specific invariance property, and somewhat weaker support for the cluster of invariance properties associated with the attribute. As the number of facets that is investigated without encountering refutations of their invariance properties increases, the degree of support for the conjecture that the measurement procedure adequately represents the attribute increases.

G studies, which do not sample randomly from a facet, but, instead, sample from a restricted subset of the universe for the facet, don't provide adequate evidence for invariance over the full facet. They do provide evidence for invariance over the subuniverse sampled, they provide some support for invariance over the facet, and they also provide some support for invariance over the universe of generalization as a whole. Such a G study provides only weak evidence for the validity of a measurement procedure, and it may require a large number of such G studies to develop a high degree of confidence in the interpretation of observed scores as valid measurements of an attribute.

The evaluation of evidence for dispositional validity is complicated further by the fact that the support for invariance over the universe of generalization that is provided by evidence for invariance over a particular facet will depend on the facet studied. If there is some reason to suspect that a particular facet may have a large effect on observed scores, evidence for invariance over that facet provides relatively strong support for invariance over the universe of generalization. For example, in evaluating the dependability of performance ratings, the variance components for raters could be substantial, while the variance component for equipment would be negligible in many cases. Therefore, an investigation of the rater facet provides a more severe test than an investigation of the equipment facet, and passing the more severe test provides stronger evidence for the overall dependability of the measurement procedure than passing the weaker test.

RANDOM SAMPLING FROM THE UNIVERSE OF ALLOWABLE OBSERVATIONS

The random sampling of observations is important for the measurement of dispositional attributes in two ways. First, a measurement procedure is defined in terms of random samples from the universe of allowable observations (D studies). Therefore, the application of the measurement procedure to any object of measurement requires the random sampling of observations from the universe of allowable observations for the object of measurement. Second, studies of the properties of measurement procedure (G studies) require random sampling from the population of objects of measurement and random sampling from the universe of generalization.

The purpose of G studies is to provide data that can be used in the design of effective measurement procedures. In particular, an important goal in designing a measurement procedure is to reduce the number of facets that must

be randomly sampled in obtaining an observed score and there are three ways to do this. First, if G studies show that all of the variance components (for the main effect and interactions) for a facet are zero, there is no need to be concerned about how this facet is sampled. Second, facets that have been standardized are not sampled in estimating universe scores, but the variance of the systematic errors for these facets must be estimated in G studies. Third, if the universe of generalization is restricted in connection with the development of theory, thus defining a new attribute, the universe of allowable observations for measurements of this new attribute will also be restricted; that is, if the new attribute does not involve generalization over a facet, the measurement procedure will not involve sampling of the facet. Such theoretical developments are also based, in part, on the results of G studies.

All of these modifications of the measurement procedure tend to decrease the number of facets which must be randomly sampled in obtaining an observed score. The only facets that need to be sampled randomly are those for which interactions with the objects of measurement are fairly large and apparently random. Efforts to obtain random samples can be concentrated on these facets, and as the number of such facets decreases, the difficulty in taking random samples from the universe of allowable observations is reduced.

One of the results of these changes is, therefore, to transfer the burden of random sampling from the measurement procedure to the G studies. If these efforts to reduce or eliminate the random errors for various facets are successful, the variance components for all of the facets in the universe of allowable observations will be small, and, therefore, the measurement procedure will have a high reliability.

To the extent that the universe of allowable observations is homogeneous in the sense that all observations in this universe yield approximately the same value of the observed score for each object of measurement, it does not matter which observation is selected from the universe of allowable observations, or how this observation is chosen. This is especially true when the random error variance is small compared to the variance of specific systematic errors. To the extent that this kind of homogeneity is attained in the universe of allowable observations, the measurement procedure is robust against violations of the random sampling assumption for the facets in the universe of allowable observations.

This assertion may seem extraordinary since statistical conclusions are never robust against violations of their sampling assumptions, but the situation being described is not one that is typically encountered in statistics. In most statistical analyses, the population is fixed and the aim of the study is to estimate some parameter for the population. In order to obtain unbiased estimates of the parameter, it is necessary to sample randomly.

In developing a measurement procedure, the situation is quite different. Here, the universe of allowable observations from which observations are to be drawn is not fixed. The goal is to make the variance in observed scores for each object of measurement as small as possible by refining the definition of the universe of allowable observations. To the extent that this goal is achieved, all observed scores in the universe of allowable observations for each object of measurement will be approximately the same, and it will not matter which observation is chosen.

For example, in measuring length, it isn't necessary to randomly sample from the universe of meter sticks, because it is known that all meter sticks give essentially the same result. Therefore, the most convenient meter stick is used. The justification for such practices is found in the empirical generalization that the variance introduced into observations by the choice of meter stick is very small compared to the variance introduced by some other factors (e.g. temperature).

GENERALIZABILITY COEFFICIENTS AS UPPER BOUNDS ON VALIDITY

The main source of difficulty with the simple version of the sampling model is that it ignores the complex sampling procedures that are actually employed by measurement procedures. In order to obtain a single point estimate of the validity, this simple model makes the unrealistic assumption that observed scores are based on random samples from the universe of generalization. In practice, this assumption does not apply whenever any facet is standardized. The explicit recognition of standardization leads to a more complex model which does not claim to provide point estimates of validity, but aims instead to provide a series of upper bounds on the validity.

Standardization changes random errors into systematic errors. The evaluation of systematic errors due to standardization requires that the sampling variability for the standardized facet be estimated independently of the other facets, and there may be many facets that are standardized in a given measurement procedure. In general, it is not practical to draw random samples from the universe of generalization and therefore all the invariance properties for an attribute cannot be evaluated in a single G study. In particular, it is usually not possible to obtain independent estimates of variance components for more than a few facets without having very large sample sizes.

A series of upper bounds is perhaps a less satisfactory result than a point estimate of the validity, but it is generally more realistic to consider the coefficient resulting from a typical G study as an upper bound on validity than as an unbiased point estimate.

THE PROCRUSTES EFFECT IN DEFINING UGs

Throughout most of this paper, it has been tacitly assumed that the universe of generalization defining an attribute is fixed; and that the task is to investigate the invariance properties implied by the attribute's definition. For purposes of exposition, these assumptions have been convenient, but, in practice, the situation is never quite this simple (see Cronbach, 1971, p.482).

The definition of the universe depends, at least in part, on the invariance properties that can be established. Initially the definition of the universe is likely to be very loose, with many facets defined vaguely by standard expressions such as, "within normal limits". Over time, the universes of conditions for various facets may be clarified, as more is learned about how various conditions of the facet influence observations. For example, if it is found that observations depend strongly on the choice of condition for a particular facet, it may be necessary to restrict the definition of the universe for that facet to a particular condition, or to a small set of conditions.

On the other hand, if an attribute which was expected to vary with different conditions of a particular kind, is found to be invariant over these conditions, the universe of generalization for the attribute may be extended to include this kind of condition as a new facet.

In most cases, decisions about whether or not to generalize over a particular class of conditions will depend, in part, on whether observations are invariant over the conditions. If the observations are not invariant over the class of conditions, including these conditions as a facet in the universe of generalization would decrease the validity of measurements of the attribute; therefore the conditions are not likely to be defined as a facet of the universe of generalization. However, if the observations are at least approximately invariant over the class of conditions, broadening the definition of the attribute to include these conditions as a facet would not decrease validity, and would increase the usefulness of the measurements.

THE STEADY STATE REQUIREMENT

Cronbach et al(1972) raise an issue which they treat as a limitation of generalizability theory:

> "Because our model treats conditions within a facet as unordered it will not deal adequately with the stability of scores that are subject to trends, or to order effects arising from the measurement process...a large contribution will be made by the development of a model for treating ordered facets..."(p. 364)

It is clearly inappropriate to consider consecutive observations as randomly sampled from a universe of generalization if these observations are known to depend systematically on time or on any other facet in the universe. However this should not be viewed as a limitation in a theory of measurement. Assuming that the theory of measurement is intended to analyze the methods rather than the substantive content of science, there is no need for it to cover functional relationships among different variables.

According to the domain sampling model proposed here, to generalize over a facet is to treat the variability of observed scores due to the sampling of the facet as error. Where the conditions of some kind are considered a facet, the observed score is interpreted as an estimate of the mean over all conditions of the facet, and the observed score is not associated with a particular condition of the facet. On the other hand, in order to recognize a relationship between observed scores and the conditions of a facet, each observed score must be associated with a particular condition of the facet, and this implies the absence of generalization over the facet. Therefore, within this model, it is inconsistent to say that a particular kind of condition should be included as a facet in the universe of generalization, and at the same time, to say that observed scores are a function of the facet.

The problems introduced into sampling models by the existence of trends can be eliminated as soon as the trend is detected; this is accomplished by restricting the universe of generalization for each observation to a fixed condition of the facet involved, and by treating the trend as an empirical law (see section VI). Undetected trends will tend to cause the variance components for the facet to be large, and therefore the examination of variance components can facilitate the detection of trends.

CONCLUDING COMMENTS

The sampling model provides a framework for considering the issues that arise naturally in the interpretation of measurements in terms of dispositions. The three types of issues that have been identified are those associated with reliability, validity, and import. Reliability indicates how well observed scores represent the universe of allowable observations, Validity indicates how well observed scores represent the universe of generalization defining an attribute, and Import indicates how well the observed score predicts other observed scores that are of interest.

Since measurable attributes in both the physical and behavioral sciences are interpreted as dispositions, these issues arise for every measurement procedure. However, the way in which these issues should be analyzed is not fixed. For convenience, most of the discussion in this paper has been in terms of variance components and generalizability coefficients, but the same points could have been made in terms of correlation coefficients. In fact, where one is interested in the relationship between the observed scores for particular conditions of a facet and universe scores, it would be natural to use correlation coefficients. In dealing with categorical data, a rather different set of indices would need to be estimated.

Although variance components seem to match the assumptions of the sampling model for dispositions especially well, the formal statistical models defining variance components should not be allowed to obscure the fundamental concerns embodied in the invariance properties. As Cronbach(1976) has observed, the technical apparatus of generalizability theory is less important than the questions suggested by the theory.

The sampling model has a number of advantages. It is formulated in terms of the fundamental statistical concept of random sampling, and the model is basically quite simple. An attribute is defined in terms of a universe of generalization, and the universe score for the attribute is simply the expected value over the universe of generalization.

If one wishes to assume that the invariance properties associated with a universe of generalization reflect some underlying structure or process associated with the attribute, one is free to do so. However the definition of the universe of generalization can also be treated as a matter of convenience or convention. The sampling model is consistent with either of these two points of view.

Although the sampling model makes few assumptions, it provides an analysis of many issues associated with the dependability of measurement. It makes it possible to give validity a straightforward interpretation, and to draw a clear distinction between reliability and validity. The conclusions that reliabilty is an upper bound on validity and that some means of improving reliability may cause validity to decrease can be easily derived from the model. Furthermore, the model provides a basis for a detailed analysis of standardization and of the resulting systematic errors. Convergent validity can be analyzed in terms of the standardization of a method facet.

As shown in section VI, the model suggests an explicit mechanism for relating the refinement of measurement procedures to the development of laws. Although the analysis of this mechanism has not been carried very far, it does begin to clarify the relationship between theory and measurement.

The problems associated with sampling models are no more serious than the problems associated with other models. Rather, these problems are noticed more clearly because the assumptions of sampling models are more clearly stated than for other models.

American Psychological Association. Standards for educational and psychological tests. (Rev. ed). Washington, D.C.: American Psychological Association, 1974.

Bechtoldt, M. P. Construct Validity: A Critique. American Psychologist, 14, 1959, 619-629.

Braithwaite, Richard B., Scientific Explanation. Cambridge: Cambridge University Press, 1953.

Brennan, R. L. Generalizability analysis: principles and procedures. ACT Technical Bulletin no. 26 Iowa City, Iowa, American College Testing Program, 1977.

Brennan, R. L. and Kane, M. T. Generalizability theory: a review. New Directions for Testing and Measurement. 1979, 33-51.

Bridgman, Percy W. The Logic of Modern Physics. New York: Macmillan, 1927.

Campbell, D. T., and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 81-105.

Campbell, D. T. & Tyler, B. B. The construct validity of work-group morale measures. Journal of Applied Psychology, 1957, 41, 91-92.

Campbell, Norman R. What is Science? London: Methuen, 1921.

Campbell, Norman R. Physics: The Elements. Cambridge: Cambridge University Press, 1920.

Cardinet, J., Tourneur, Y., and Allal, L. The symmetry of generalizability theory: applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-134.

Carnap, R. Testability and meaning. In H. Feigl and M. Brodbeck (Eds.). Readings in the Philosophy of Science. New York, Appleton-Century-Crofts, 1953, 47-92.

Corben, H. & Stehl, P. Classical Mechanics, Second Edition. John Wiley & Sons, Inc., 1960, New York

Cornfield, J. & Tukey, J. W. Average Values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L. J. On the design of educational measures. In D.N.M. de Gruijter and L. J. T. van der Kamp (Eds.). Advances in Psychological and Educational Measurement. New York, John Wiley and Sons, 1976.

Cronbach, L. J., Test Validation. in R. L. Thorndike, ed. Educational Measurement 1971, pp. 443-507.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972

Cronbach, L. J., & Gleser, G. C. The signal/noise ratio in the comparison of reliability coefficients. Educational and Psychological Measurement, 1964, 24, 467-480.

Cronbach, L. J., and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 52, 1955, 281-302.

Ebel, R. L. Explorations in reliability theory. Contemporary Psychology, 1974, 19, 81-83.

Ennis, R. H. Operational Definitions. In H. S. Broudy, R. H. Ennis, and L. I. Krimerman. (Eds.) Philosophy of Educational Research. New York, John Wiley and Sons, 1973, 650-669.

Fiske, D. W. Can a personality construct be validated empirically? Psychological Bulletin, 1973, 80, 89-92.

Frank, P. Philosophical interpretations and misinterpretations of the theory of relativity. In H. Feigl and M. Brodbeck (Eds.). Readings in the Philosophy of Science. New York, Appleton-Century-Crofts, 1953, 213-231.

Frank, Philipp. Philosophy of Science. Englewood Cliffs, N.J.: Prentice Hall, 1957.

Gillmore, G. M. An introduction to generalizability theory as a contributor to evaluation research EAC Report No. 79-14, Seattle, Educational Assessment Center, University of Washington, 1979.

Hempel, C. G., Operationism, observation, and theoretical terms. In A. Danto and S. Morgenbesser (Eds.). Philosophy of science. New York, New American Library, Inc., 1960, 101-120.

Hempel, Carl G. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. Glencoe, Ill.: Free Press, 1965.

Hempel, C. S., Fundamentals of Concept Formation in Empirical Science. Chicago: The University of Chicago Press, 1952.

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Cal.: Wadsworth, 1968.

Kuhn, Thomas S. The Structure of Scientific Revolutions, Second Edition. International Encyclopedia of Unified Science, Vol. 2 No.2 Chicago, University of Chicago Press, 1970.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston, Houton Mifflin, 1953.

Loevinger, L. Person and population as psychometric concepts. Psychological Review, 72, 1965, 143-155.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

MacCorquodale, Kenneth and Meehl, Paul E. On a Distinction between Hypothetical Constructs and Intervening Variables. Psychological Review, 55, 1948, 95-107.

McDonald, R. P. Generalizability in factorable domains: "domain validity and generalizability." Educational and Psychological Measurement, 38, 1978, 75-79.

Meehl, P. E., On the circularity of the law of effect. Psychological Bulletin, 1950, 47.

Nunnally, J. C. Psychometric Theory. New York, McGraw Hill, 1967.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Pap, A., The a priori in physical theories. New York, Russell & Russell, 1946.

Physical Science Study Committee, College Physics. Uri Haber-Schain (ed.), Raytheon Education Company, 1968.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J., Prentice-Hall, 1978.

Popper, Karl R. The Logic of Scientific Discovery. New York, Harper and Row, 1968.

Popper, Karl R. Conjecture and Refutations: The Growth of Scientific Knowledge. New York, Harper and Row, 1965.

Rajaratnam, N., Cronbach, L. J., and Gleser, G. C. Generalizability of Stratified-Parallel Tests. Psychometrika, 1965, 30, 39-56.

Rozeboom, W. W. Foundations of the theory of prediction. Dorsey Press, Homewood, Ill., 1966.

Smith, P. L. Sampling Errors of Variance Components in Small Sample Multifacet Generalizability Studies. Journal of Educational Statistics, 3, 1978, 319-346.

Suppes, Patrick, The Structure of Theories and the Analysis of Data. In W. F. Suppe (Ed.) The Structure of Scientific Theories. Urbana, Ill.: The University of Illinois Press, 1974.

Suppes, P., and Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.). Handbook of Mathematical Psychology, Vol. I, New York, John Wiley and Sons, 1963.

Torgerson, W. S. Theory and methods of scaling. New York, John Wiley & Sons, 1958.

Toulmin, Stephen. The Philosophy of Science. London: Hutchinson's Universal Library, 1953.

Tryon, R. C. Reliability and behavior domain validity; reformulation and historical critique. Psychological Bulletin, 54, 1957, 229-249.

Winer, D. J. Statistical prinicples in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Wright, B. D. & Douglas, G. A. Best Procedures for sample-free item analysis. Applied Psychological Measurement, 1, 1977, 281-295.